欧姆社学习漫画

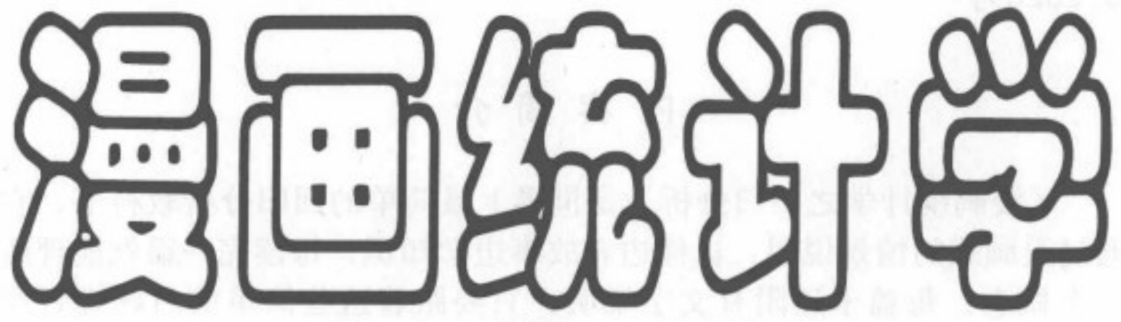
漫画统计学之回归分析

〔日〕高桥 信/著
〔日〕Inoue Iroha/漫画绘制
〔日〕株式会社TREND-PRO/漫画制作
张仲桓/译



科学出版社
www.sciencep.com

欧姆社学习漫画



之回归分析

〔日〕高桥 信 著
〔日〕Inoue Iroha 漫画绘制
〔日〕株式会社TREND-PRO 漫画制作
张仲桓 译

科学出版社
北京

图字：01-2009-2323号

内容简介

《漫画统计学之回归分析》是世界上最简单的回归分析教科书，它通过漫画式的情景说明，让你边看故事边学知识，每读完一篇就能理解一个概念，每篇末还附有文字说明，只要跟着这些简单的习题进行操练，你就能在最短时间内成为回归分析达人！

有趣的故事情节、时尚的漫画人物造型、细致的内容讲解定能给你留下深刻的印象，让你看过忘不了。通过这种轻松的阅读学习，读者可以掌握回归分析的基本知识。本书也可以作为广大青少年学习、掌握统计学中回归分析知识的读本。

图书在版编目（CIP）数据

漫画统计学之回归分析/（日）高桥 信著；（日）Inoue Iroha漫画绘制；（日）株式会社TPREND-PRO漫画制作；张仲桓译.—北京：科学出版社，2009

（欧姆社学习漫画）

ISBN 978-7-03-025006-3

Ⅰ.漫…　Ⅱ.①高…②I…③株…④张…　Ⅲ.回归分析-通俗读物　Ⅳ.O212.1-49

中国版本图书馆CIP数据核字（2009）第118728号

责任编辑：唐 璐 赵丽艳 / 责任制作：董立颖 魏 谨
责任印制：赵德静 / 封面制作：铭轩堂

北京东方科龙图文有限公司　制作
http://www.okbook.com.cn

科学出版社　出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com

北京天时彩色印刷有限公司　印刷
科学出版社发行　各地新华书店经销

*

2009年8月第　一　版　开本：787×1092　1/16
2009年8月第一次印刷　印张：14
印数：1—5 000　　字数：214 000

定价：29.80元

（如有印装质量问题，我社负责调换）

✿前　言✿

本书介绍了有关回归分析、重回归分析以及Logistic回归分析的实用统计学知识。

回归分析和重回归分析可以解决很多实际生活中的问题，例如：

- 通过“最高气温”来预测“冰红茶的销售量”。
- 通过“店铺面积”和“距离最近车站的距离”来预测“新分店备选店铺的月营业额”，并进行数值预测。

Logistic回归分析可以解决的问题例如：

- 通过“吸烟量”和“饮酒量”来预测“癌症的患病率”的概率分析。

本书的使用对象包括：

- 阅读完《漫画统计学》或者具备同等程度以上统计学知识的读者。
- 需要进行“数值预测”和“概率预测”的读者。

本书由以下4章组成：

- 第1章　基础知识
- 第2章　回归分析
- 第3章　重回归分析
- 第4章　Logistic回归分析

各章又包括：

- 漫画部分
- 对漫画部分进行补充的文字说明

第1章所讲的内容，是学习第2章及以后内容时所必备的基础知识，像微分、矩阵等等，可能是大多数读者在高中的学习过程中就已经学过的内容。“如果第1章的内容都没办法理解，还怎么阅读第2章以后的内容啊”——读者大可不必心存这种不安。在阅读的过程中，您会逐渐地体会到“原来对数就是这个意思啊！”“微分就是这样的运算啊，没错、没错，我想起来了！”不过，对于那些“完全忘记了，看了也想不起来”、“因为我是文科生，几乎没学过”的读者来说，如果不花些功夫去理解第1章的内容，就想读懂本书的主体，也就是第2章以后的内容，可能会十分辛苦。

本书中的计算过程记录得相当详细，数学基础好的读者只需仔细地看一遍即可。

数学基础稍差的读者，则要用心揣摩、多加思考。即便是那些觉得意思不太明白、计算起来也困难的读者，也请按照书中的步骤把解求出来，这样做起码可以掌握大致的计算流程。读者没必要强迫自己一次就理解，要耐心地坚持读到最后。不过在阅读过程中，请您一定要全神贯注。

在阅读中，如果存在读者自己的计算结果和书中的计算结果不一致的情况，这可能是对数据进行四舍五入的原因。如果因此给读者带来不便，还请各位读者多多谅解。

能够有这次执笔的机会，我要感谢株式会社欧姆社开发局的诸位。感谢将我的原稿制成漫画的株式会社TREND-PRO的诸位。感谢负责脚本创作的re_akino先生，以及负责绘画的Inoue Iroha先生。另外，立教大学社会学系的酒折文武先生为之前的作品提出了诸多宝贵建议，在此深表谢意。

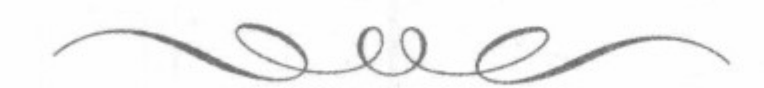
高桥　信

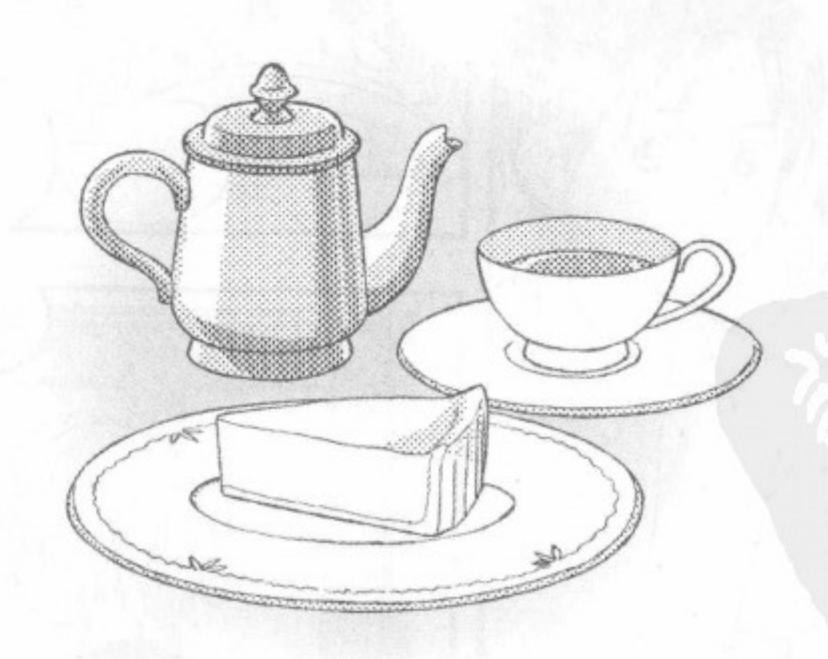


目 录

◆序　章◆

欢迎光临诺伦茶餐厅





真不错啊！
这里的红茶和蛋糕。

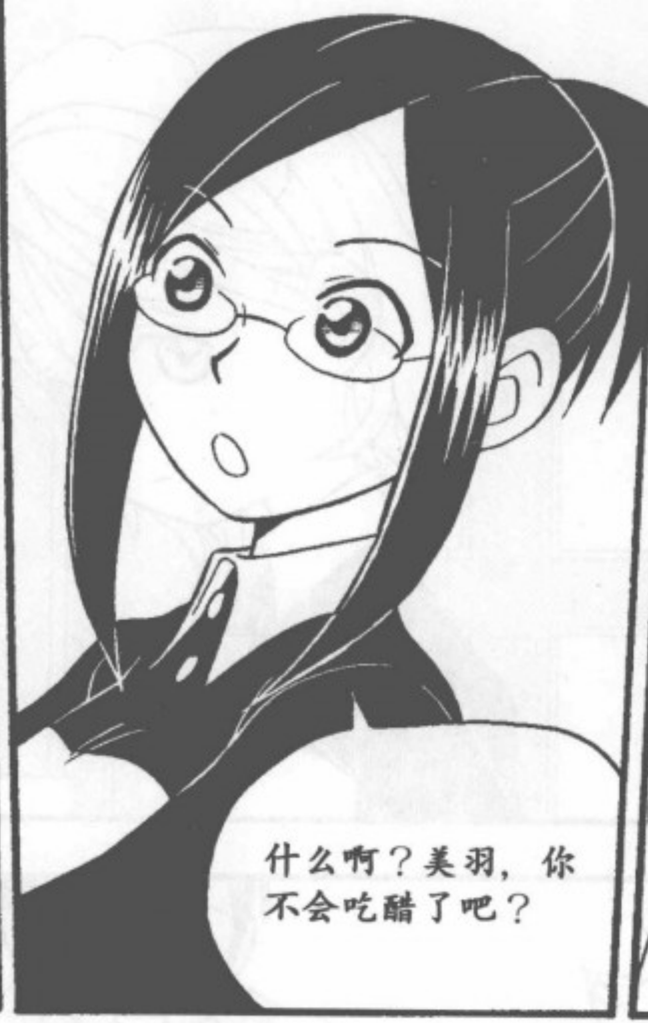

谢谢！
这可是新口味哟！


今天怎么样？
还是在学习吗？

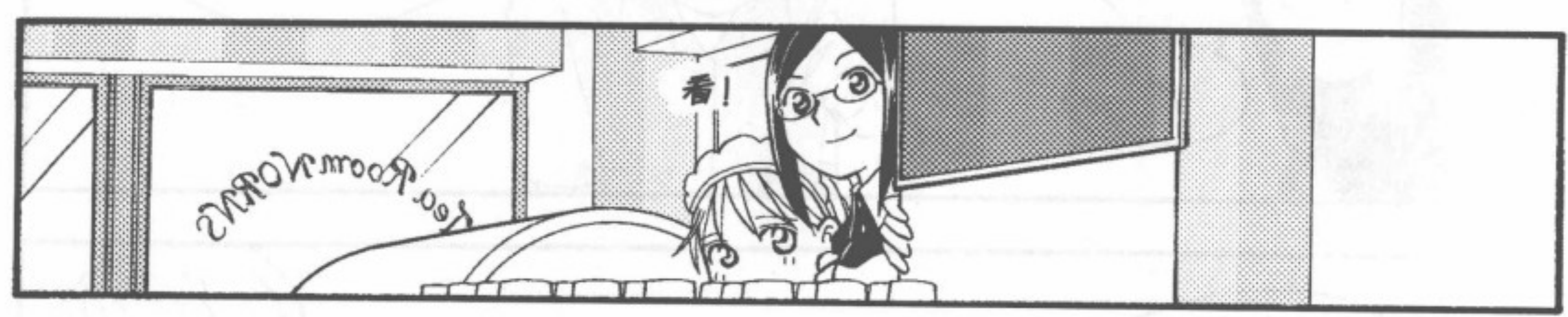

嗯！
是啊。


真不错！
くすっ
呵！
请慢用！
咦？


真、真亲切啊，
难道……
理纱前辈……


什么啊？美羽，你
不会吃醋了吧？


没，没有啊！
是吧？
是吧？


看！
Tea Room NORNS


真是个勤奋用功的
男生啊……

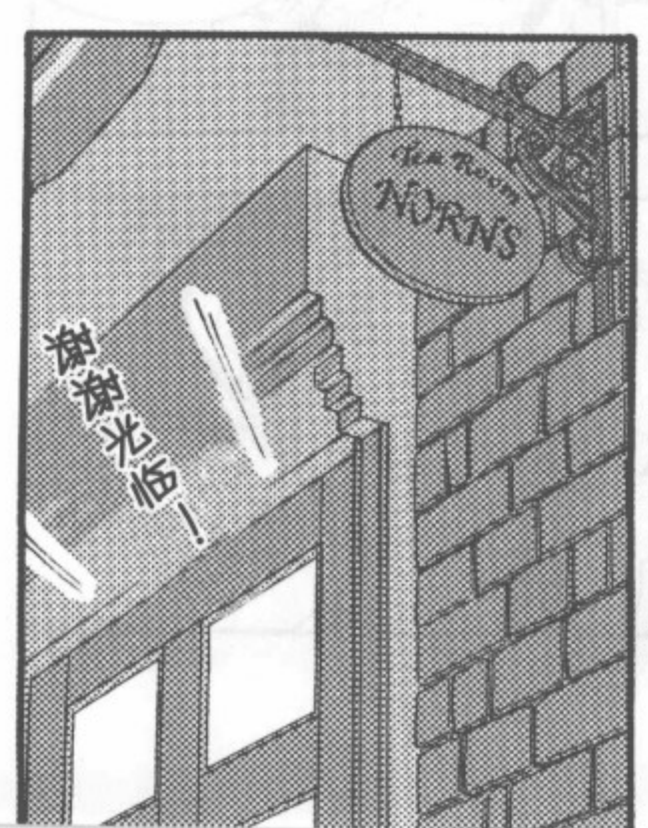

总是在看那本貌似
很难的数学书。
等一下！
我们不是“经济
学系”的吗？


要是你向
他请教的
话……
怎么说
理纱前辈也是优等
生啊！而我就完全
不……


那，那怎么行！
连名字都不知道
呢！
所以才要找机
会说话啊……
要是那个人再也
不来诺伦……
咔……
啊，
好漂亮啊！
好可
爱呀！
啊！


您好！
您好！
欢迎光临诺伦！


两位，
请随便坐！

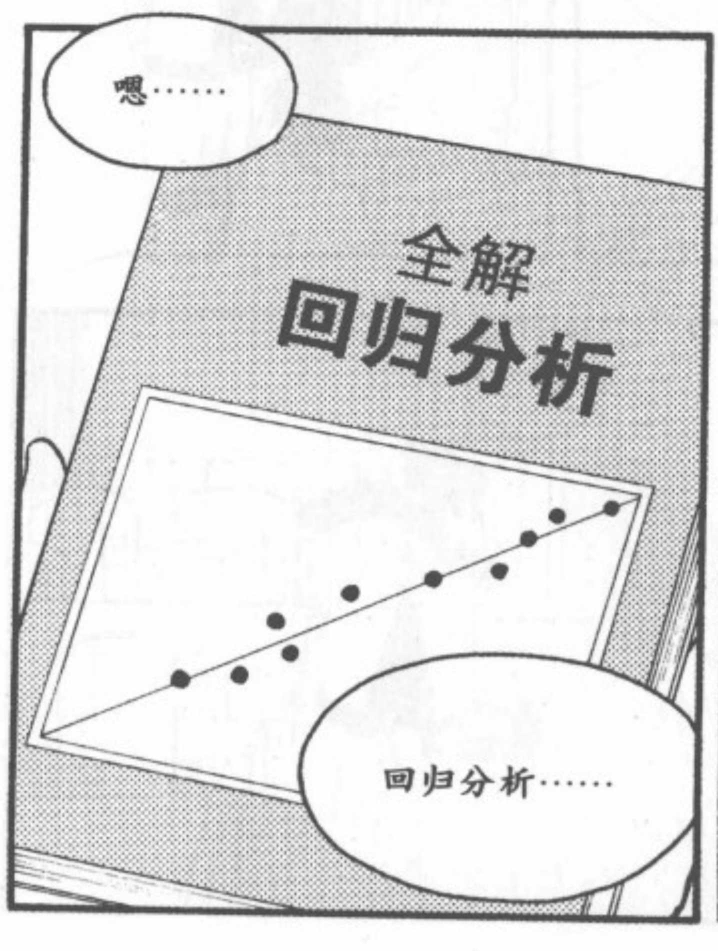

NORNS
谢谢光临！


啊！
今天……
看到了那个人的
脸……


啊！那家伙把书落下了。
是落下了吗？
嗯！


ぎっぎっ
请，请让我看一下！


瞧！
ドキッ


嗯……
全解
回归分析
回归分析……


这是一种统计学的分析方法。
哦？
抱歉，我不太了解……


美羽，是不是每天都收看电视里的天气预报啊？
是，每天都看，可是……



比方说，在诺伦我们将每天的“最高气温”和“冰红茶的销售量”都记录下来。

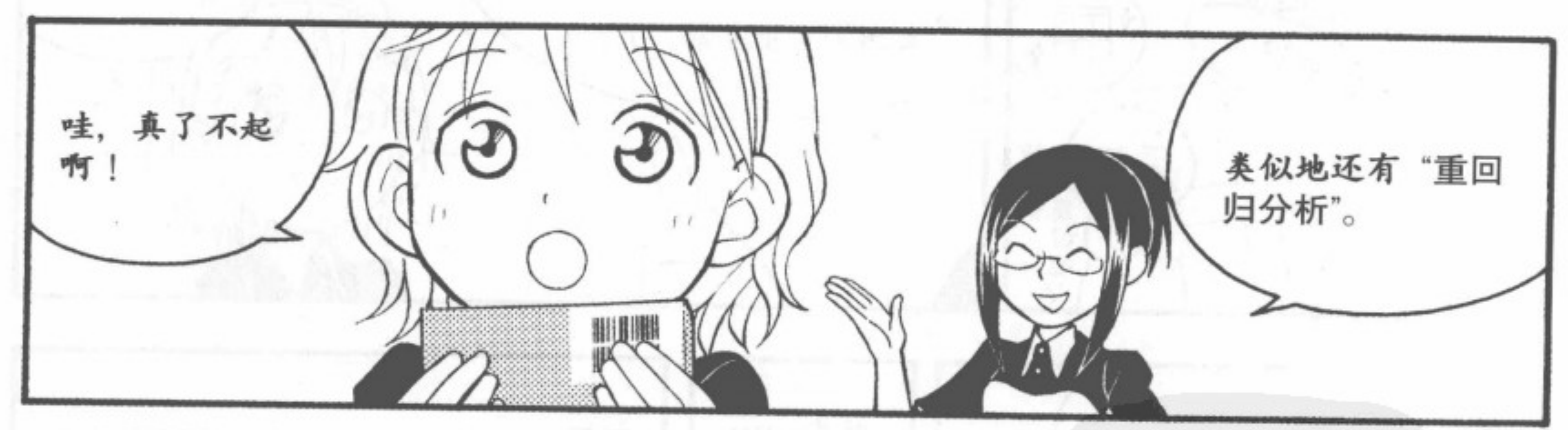


回归分析

今天的最高气温是27℃。

今天会卖出65杯冰红茶。

如果做“回归分析”，我们就可以通过最高气温来预测冰红茶的销售量了！



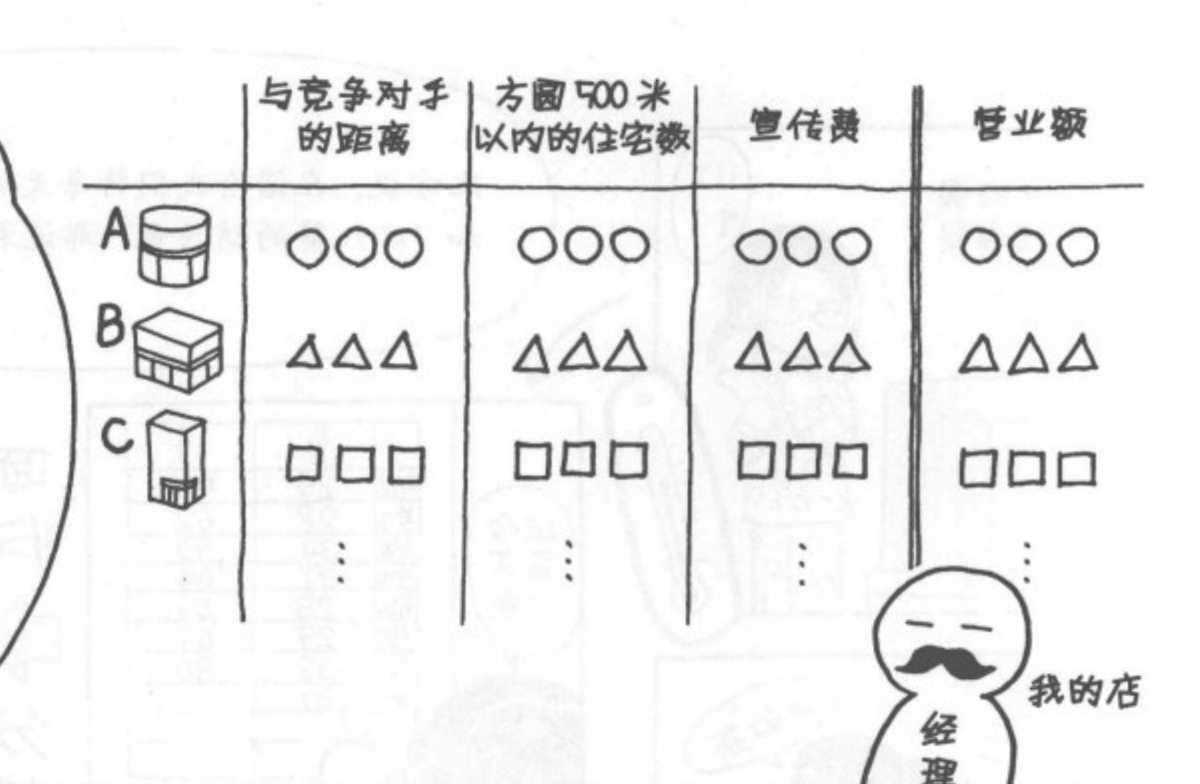




我来给你举个例子吧。某个连锁店的经理一定会掌握现有店铺的情况：

•与竞争对手的距离
•方圆 500 米以内的住宅数
•宣传费

等等，类似的数据对吧？





如果使用重回归分析的话，就能够通过

•与竞争对手的距离
•方圆 500 米以内的住宅数
•宣传费

来预测"营业额"了。

分店！

那会很简单呀！









请您教我回归分析吧！！

嗯，

拜托了！

……

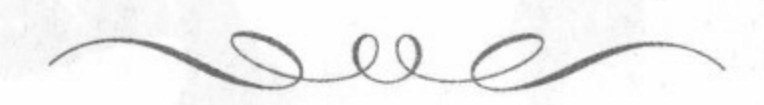
好吧！

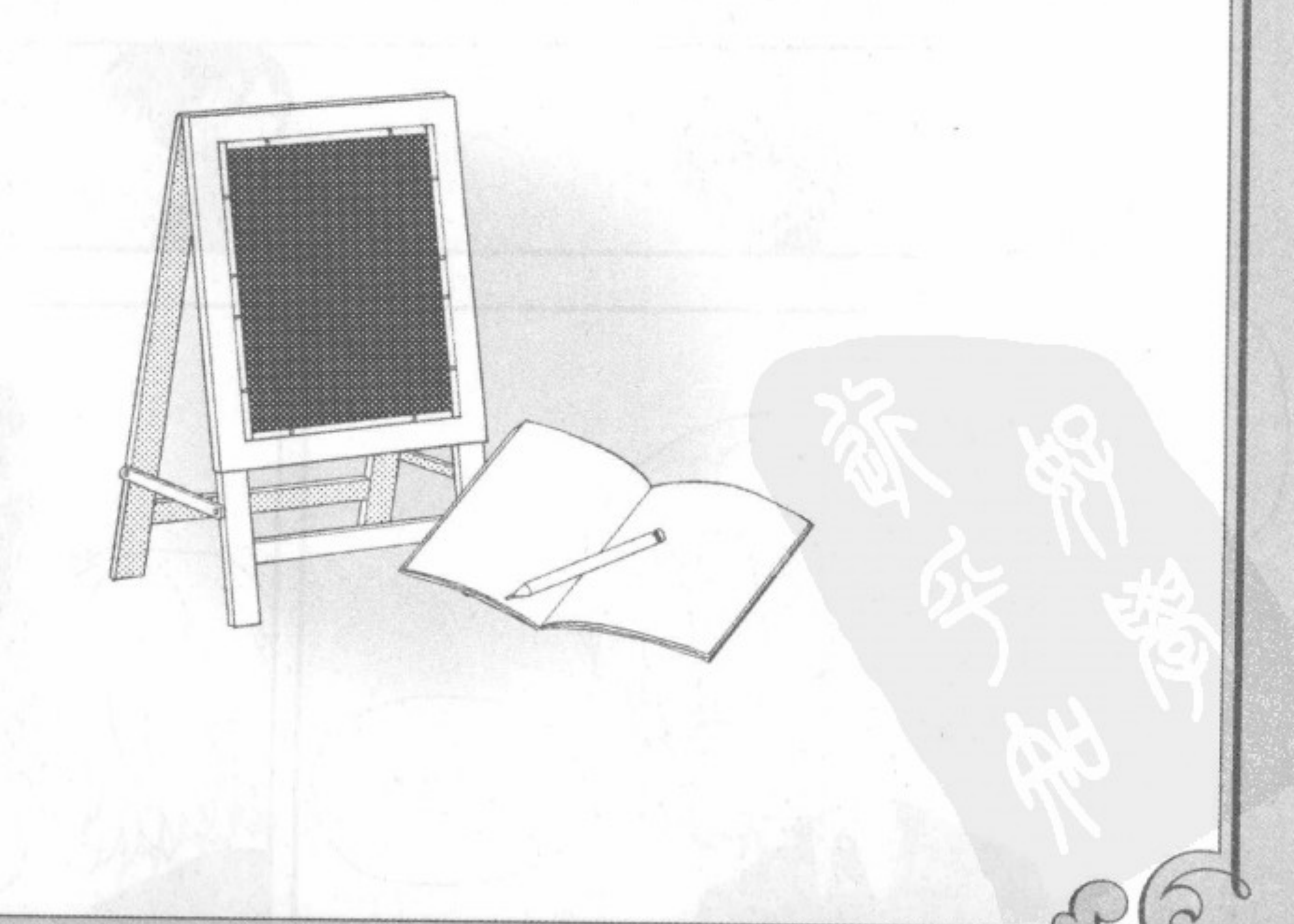
哈，
真的吗！？


真的不想把那本书直接还回去了吗？
啊……可是不放在店里也不行啊……
会给你想办法的！
啊？
ポー
可，可以吗？
放心吧！
加油！
这样的话……
就可以传达我的思念了！
好的！

◆第1章◆

基础知识





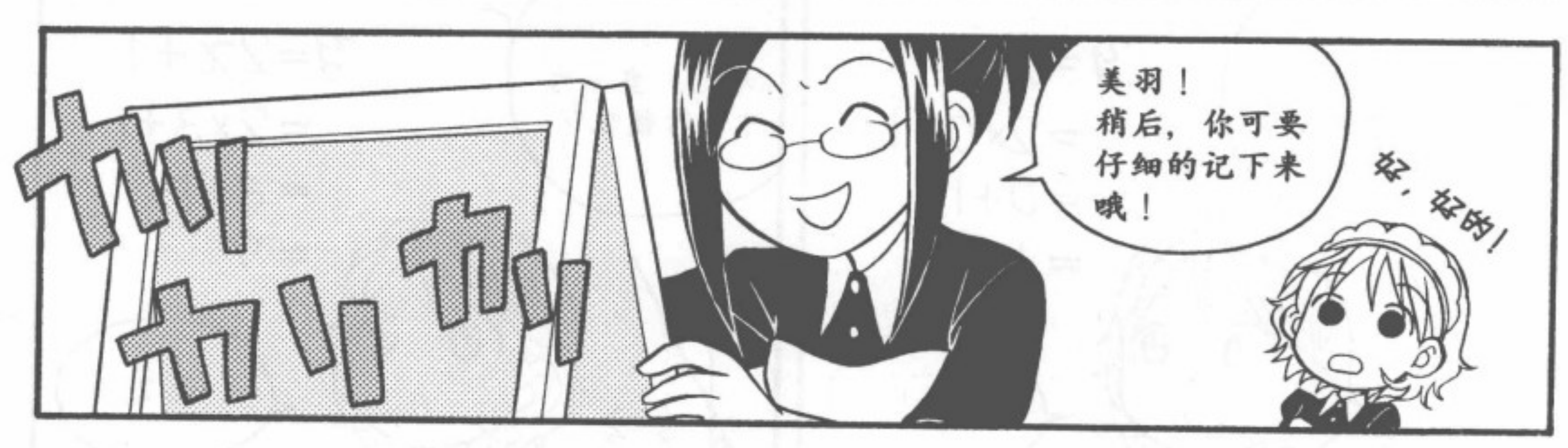
1. 书写规则



噗，噗……
老板也回去了，今天的活干完啦！
理纱前辈！请您赶快教我吧！
啊！？
现在就开始吗？
真是的……如果面对他的时候，也能这么直接的话……
是，是啊……
啊，对不起！


来吧！
不过如果我们马上学习回归分析的话会有些困难，所以今天就从基础知识开始学吧！
拜，拜托您了！


首先，我们来学习数学中的书写规则。
今日推荐
价目板……


カッ
カッ
カッ
美羽！
稍后，你可要仔细的记下来哦！
好，好的！


书写规则
$x \times x \times x = x^3$
$x \times x = x^2$
$x = x^1$
$1 = x^0$
$\frac{1}{x} = x^{-1}$
$\frac{1}{x^2} = x^{-2}$
$\frac{1}{x^3} = x^{-3}$
这个是“规则”，所以不用问为什么，只要按照它做就可以了。
是！

✿ 2. 反函数 ✿



y=2x+1
接下来，我以一次函数
y=2x+1 为例，
讲解一下反函数。



比如，当 x 为 0
时 y 的值是多少？
y=2x+1
=2×0+1
=0+1
=1
是 1！



那么，当 x 为
3 的时候呢？
y=2x+1
=2×3+1
=6+1
=7
是 7。



这样的运算很
明显，没什么
特别的。
当 x 的值为"某个"确定
的值的时候，y 的值也就
相应地确定了，对吧？
嗯！



如此说来可以把 x 看成
"主人"，y 看成"佣人"。
我渴了！
给您预备
了橙汁。
2 的 3 次
方是？
是 8。
主人
佣人







将主人和女佣的关系颠倒过来……

“女佣感谢日”来啦！

8是2的？

3次方。

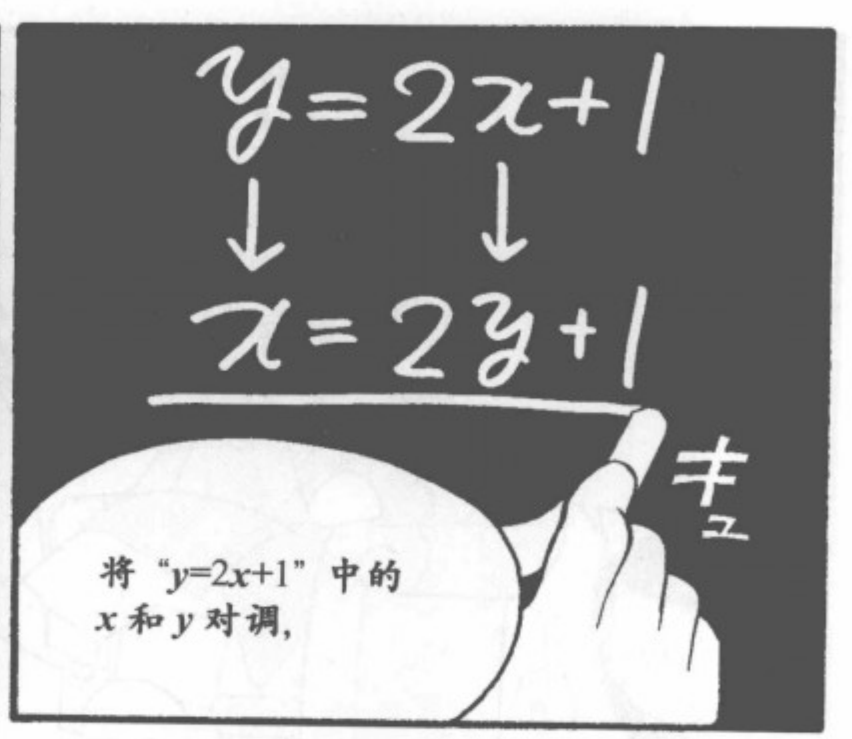
以下犯上！

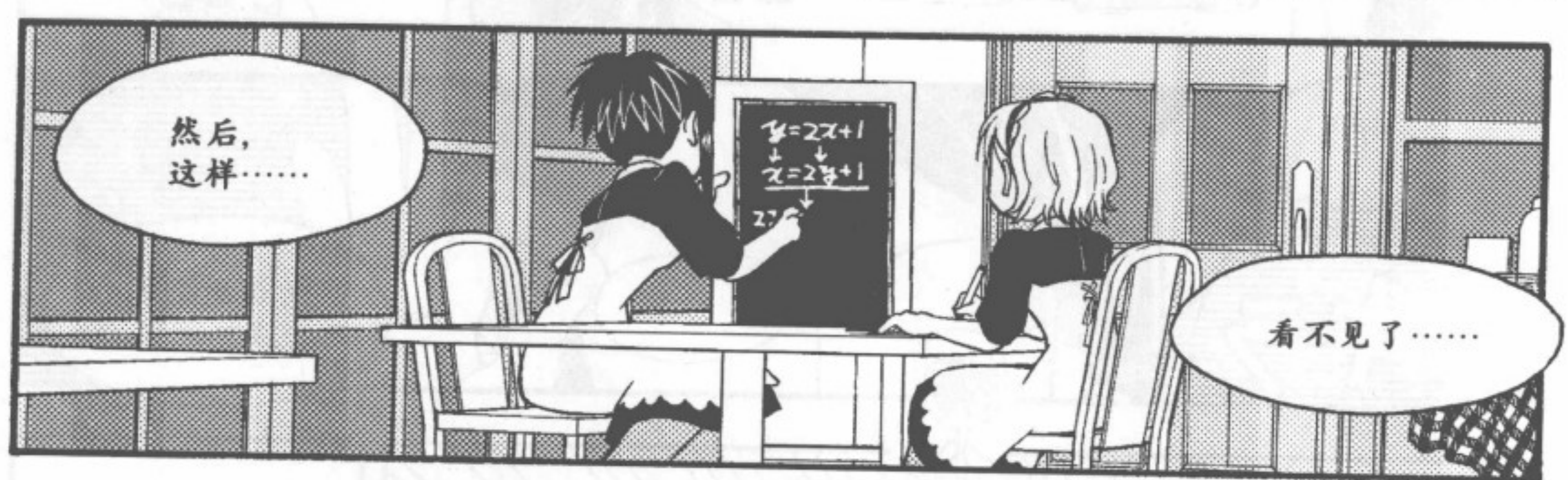
但还没有推翻！

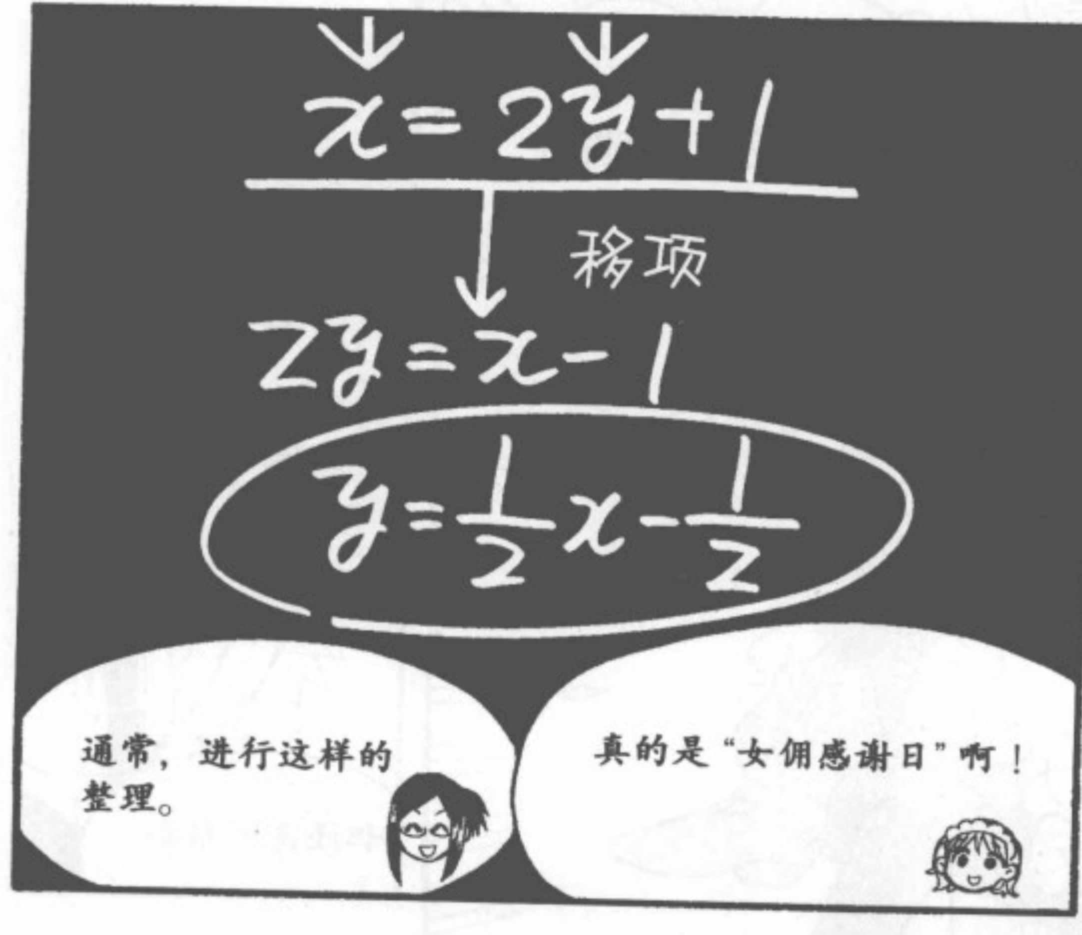
回到原来的话题，
$y=2x+1$ 的反函数就是……

$y=2x+1$

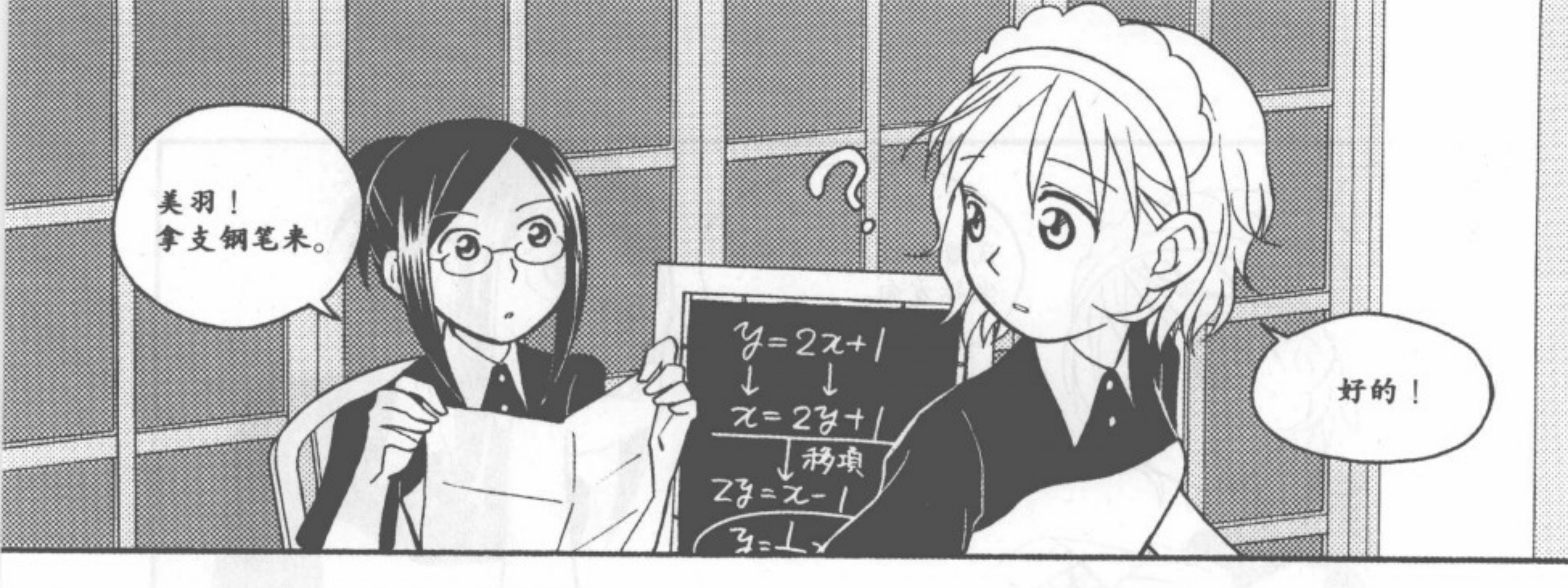
コン
コン



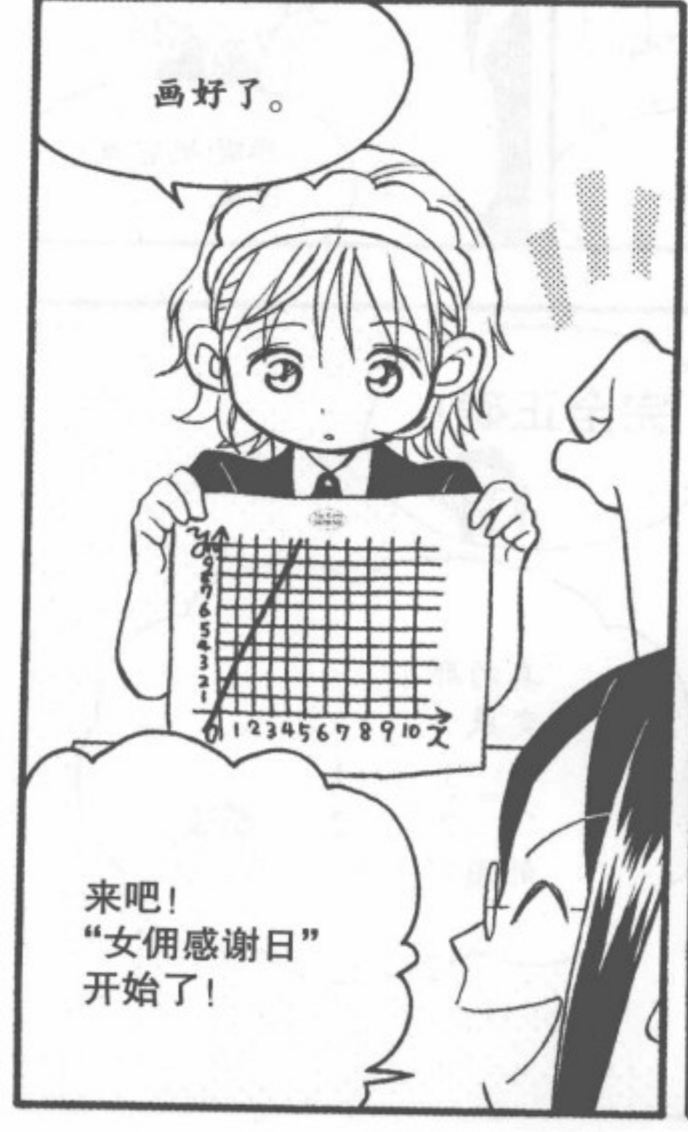


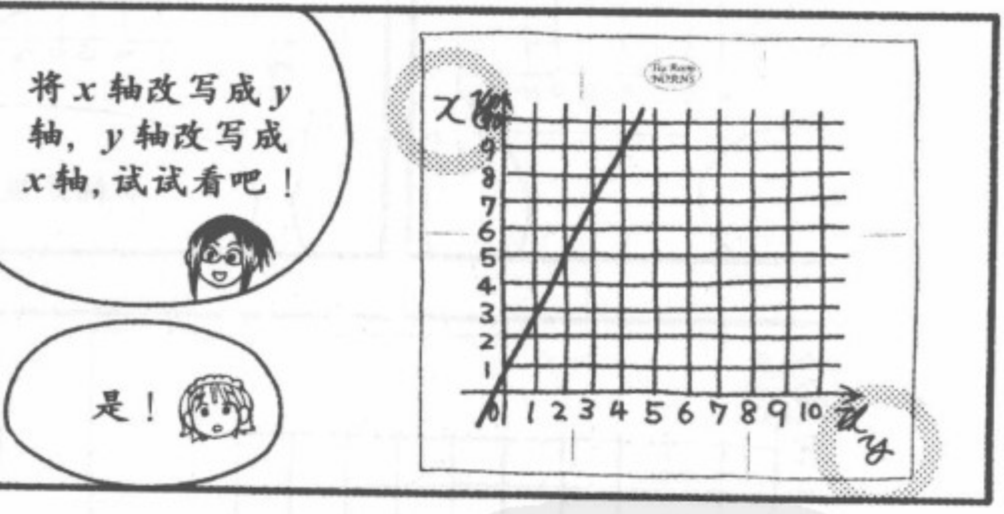


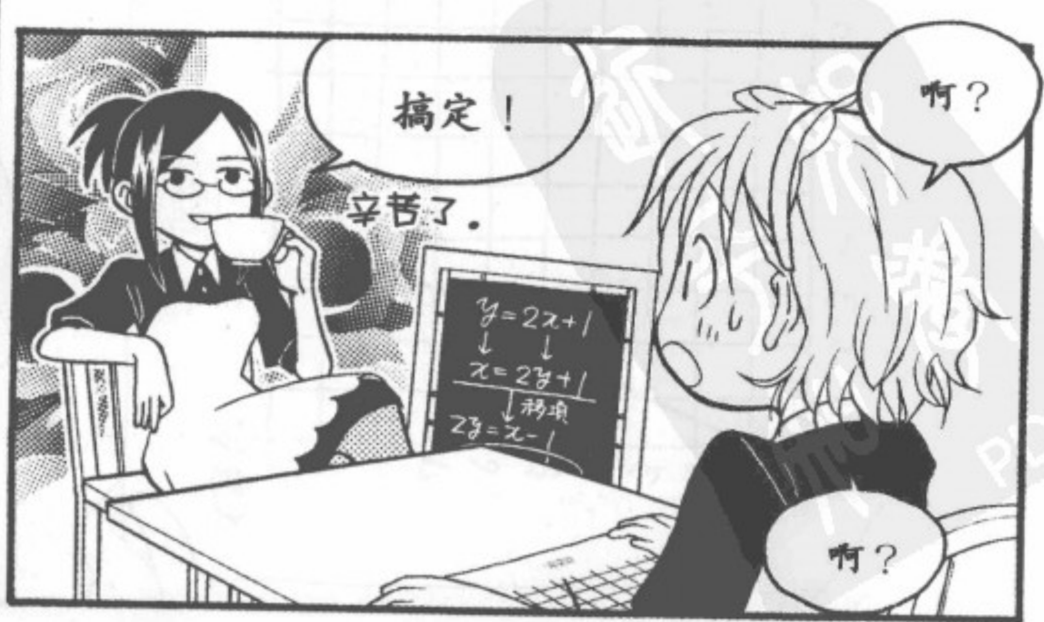



美羽！
拿支钢笔来。
好的！

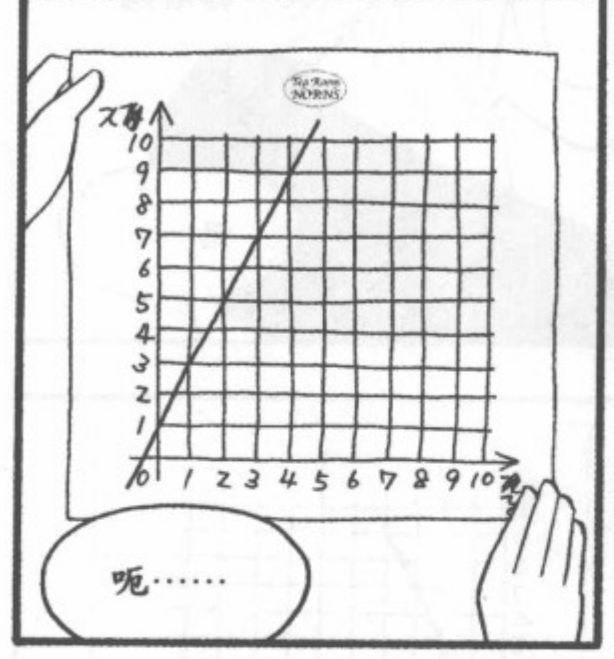

首先，画一下
y=2x+1 的图像。
好难画……
嗯……

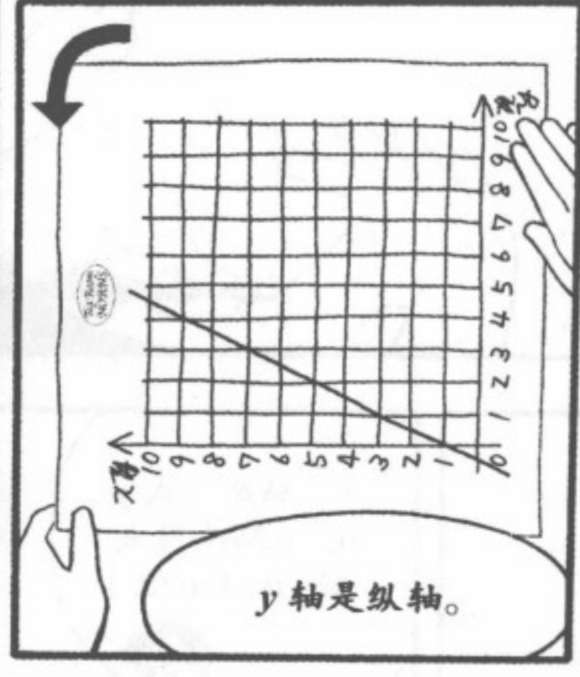

画好了。
来吧！
“女佣感谢日”
开始了！


将x轴改写成y轴，y轴改写成x轴，试试看吧！
是！

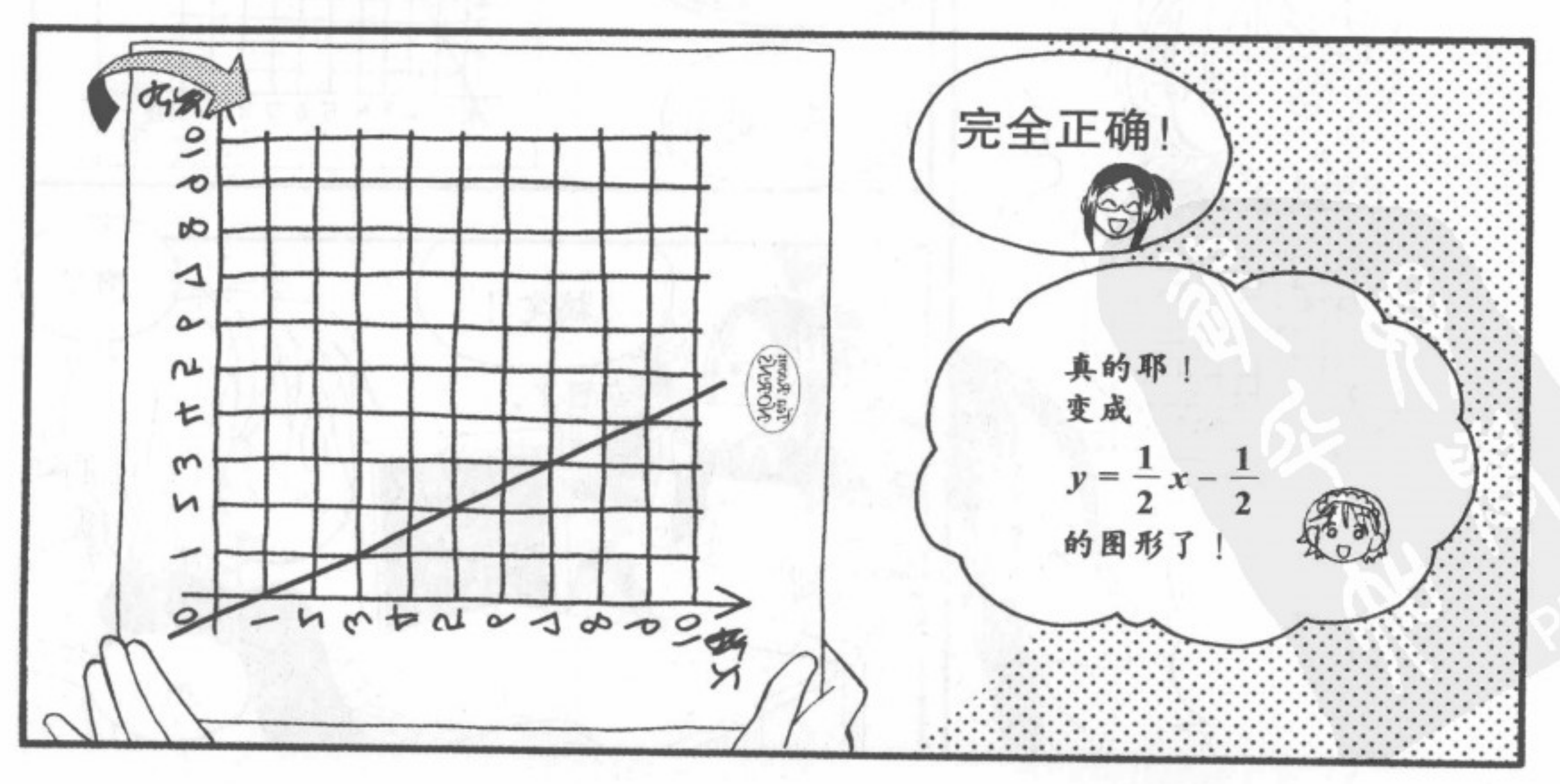

搞定！
辛苦了。
啊？
啊？


……
试着变成我们常见的坐标方向。

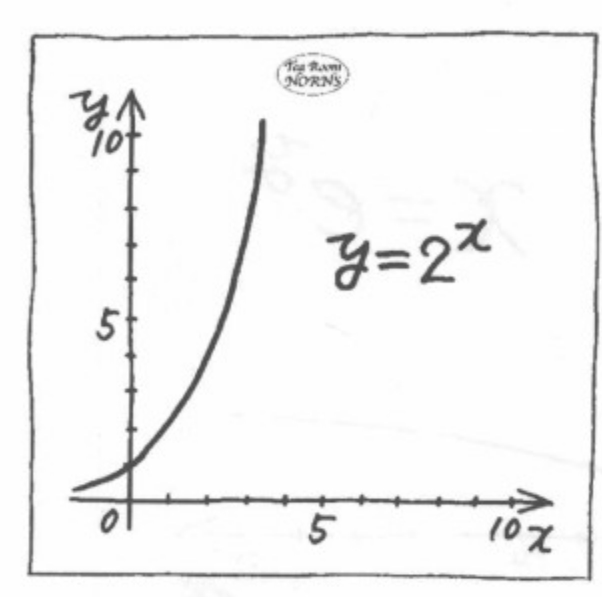

呃……

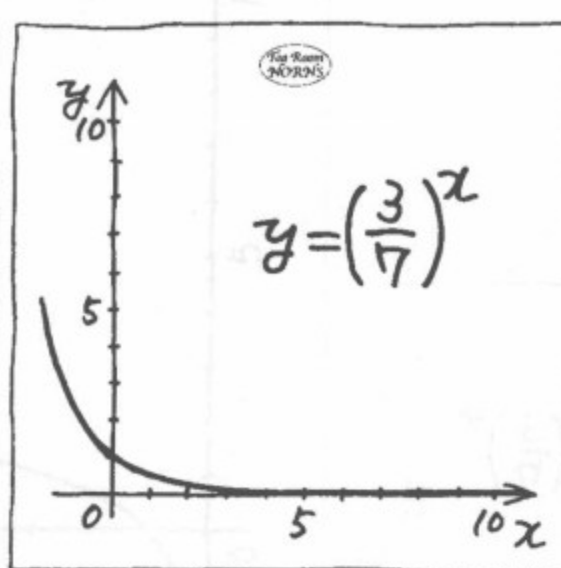

y 轴是纵轴。

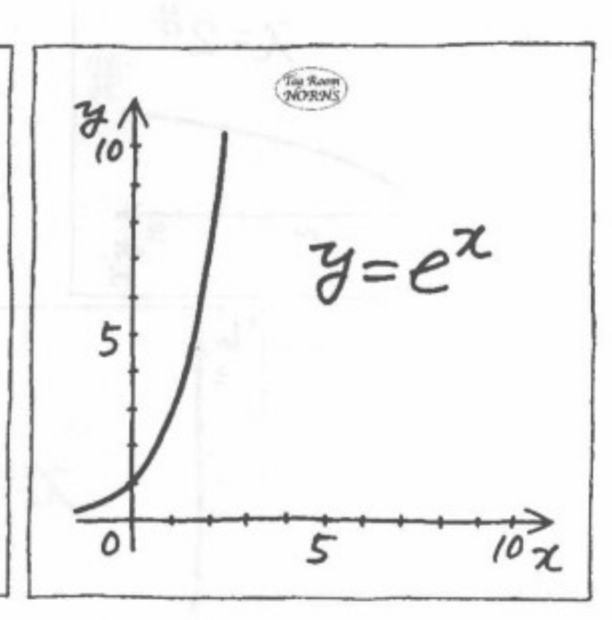

这样的话，x 轴方向不就左右相反了吗？
嗯……
那就把它反过来……


完全正确！
真的耶！变成 $y=\frac{1}{2}x-\frac{1}{2}$ 的图形了！

✿ 3. 指数函数与自然对数函数 ✿





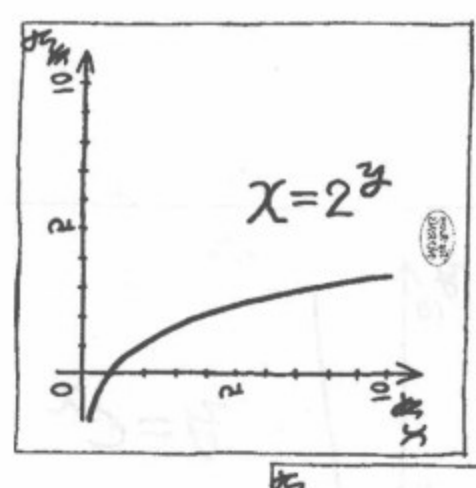


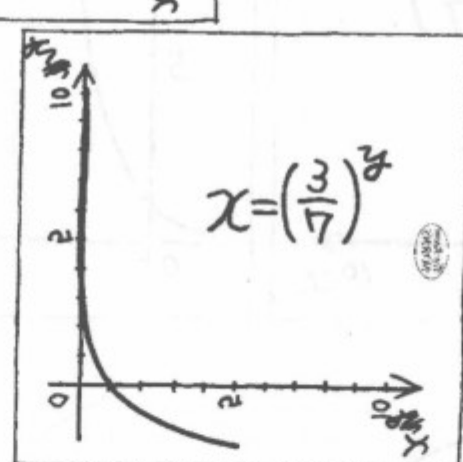


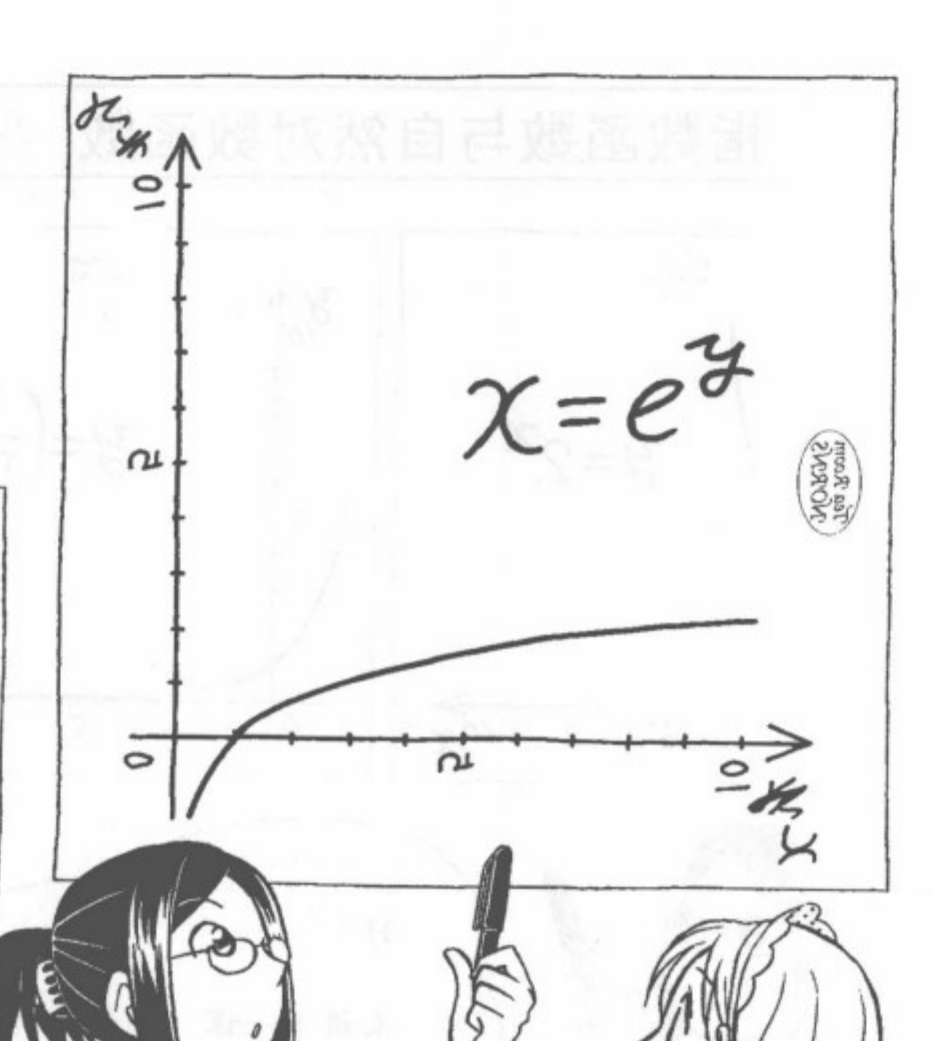












特别地，我们将 $y=e^x$ 的反函数 $x=e^y$ 称为“自然对数函数”。

噢！

但是，由于“$x=e^y$”的形式不易理解，所以我们就把 $x=e^y$ 写成“$y=\log_e x$”或“$y=\log x$”的形式。

$y=e^x$

反函数

$x=e^y$

⇕

$y=\log x$

✿ 4. 指数函数与对数函数的性质





性质 1 $(e^a)^b$ 和 e^{ab} 是等价的。

为了便于理解这条性质，我们以 $\begin{cases} a=3 \\ b=5 \end{cases}$ 为例，来具体说明，$(e^3)^5$ 和 $e^{3\times5}$ 是等价的。

让我们一起算算吧！

$$(e^3)^5 = \underbrace{e^3 \times \cdots \times e^3}_{5} = \underbrace{(e \times e \times e) \times \cdots \times (e \times e \times e)}_{5} = \underbrace{e \times \cdots \times e}_{15} = \underbrace{e \times \cdots \times e}_{3\times5} = e^{3\times5}$$



性质 2 $\frac{e^a}{e^b}$ 和 e^{a-b} 是等价的。

为了便于理解这条性质，我们以 $\begin{cases} a=3 \\ b=5 \end{cases}$ 为例，来具体说明，$\frac{e^3}{e^5}$ 和 e^{3-5} 是等价的。

让我们一起算算吧！

$$\frac{e^3}{e^5} = \frac{e \times e \times e}{e \times e \times e \times e \times e} = \frac{\not{e} \times \not{e} \times \not{e}}{e \times e \times \not{e} \times \not{e} \times \not{e}} = \frac{1}{e^2} = e^{-2} = e^{3-5}$$

性质 3　a 和 $\log(e^a)$ 是等价的！



为了便于理解这条性质，我们以 $a=3$ 为例，来具体说明，3 和 $\log(e^3)$ 是等价的。

让我们一起算算吧！

在第 20 页，我们曾讲过，$y=\log x$ 与 $x=e^y$ 是同一个意思。这就是说，如果设 $\log(e^3)$ 为 L，那么，$L=\log(e^3)$，又因为 $L=\log(e^3)$ 与 $e^3=e^L$ 等价，我们就可以把 $e^3=e^L$ 改写成下面的形式，

$$e^3=e^L$$

$$3=L$$

由于 $L=\log(e^3)$，所以等式 $3=\log(e^3)$ 成立。

性质 4　$\log(a^b)$ 和 $b\times(\log a)$ 是等价的！



为了便于理解这条性质，我们以 $\begin{cases}a=3\\b=5\end{cases}$ 为例，来具体说明，$\log(3^5)$ 和 $5\times\log 3$ 是等价的。

让我们一起算算吧！

设 $L=\log 3$，而 $L=\log 3$ 与 $3=e^L$ 是等价的。我们就可以把 $3=e^L$ 改写成如下形式。

$$3=e^L$$

$$3^5=(e^L)^5$$ 两边同取 5 次方

$$3^5=e^{L\times 5}$$ 根据性质 1

$$3^5=e^{5\times L}$$

$$\log(3^5)=\log(e^{5\times L})$$

$$\log(3^5)=5\times L$$ 根据性质 3

由于 $L=\log 3$, 所以等式 $\log(3^5)=5\times(\log 3)$ 成立。

性质 5　$\log a+\log b$ 和 $\log(a\times b)$ 是等价的！



为了便于理解这条性质，我们以 $\begin{cases}a=3\\b=5\end{cases}$ 为例，来具体说明，$\log3+\log5$ 与 $\log(3\times5)$ 是等价的。

让我们一起算算吧！

设 $\begin{cases}L=\log3\\M=\log5\\N=\log(3\times5)\end{cases}$，则 $\begin{cases}L=\log3\\M=\log5\\N=\log(3\times5)\end{cases}$ 与 $\begin{cases}3=e^L\\5=e^M\\3\times5=e^N\end{cases}$ 是等价的。

我们可以把 $e^L\times e^M=3\times5$ 改写成

$$e^L\times e^M=\underbrace{e\times\cdots\times e}_{L}\times\underbrace{e\times\cdots\times e}_{M}=\underbrace{e\times\cdots\times e}_{L+M}=e^{L+M}=3\times5$$

所以 $e^L\times e^M=3\times5=e^N$ 成立。

由于 $L+M=N$，

所以 $\log3+\log5=\log(3\times5)$ 成立。

这里我们总结一下刚刚讲过的性质。

性质 1	$(e^a)^b$ 和 $e^{a\times b}$ 等价。
性质 2	$\frac{e^a}{e^b}$ 和 e^{a-b} 等价。
性质 3	a 和 $\log(e^a)$ 等价。
性质 4	$\log(e^b)$ 和 $b\times(\log a)$ 等价。
性质 5	$\log a+\log b$ 和 $\log(a\times b)$ 等价。

顺便说一下，在这些性质中的 e，并不是像其他的 2、$\frac{3}{7}$ 这样的数，可以随意替换的哟！



✿5. 微　分✿



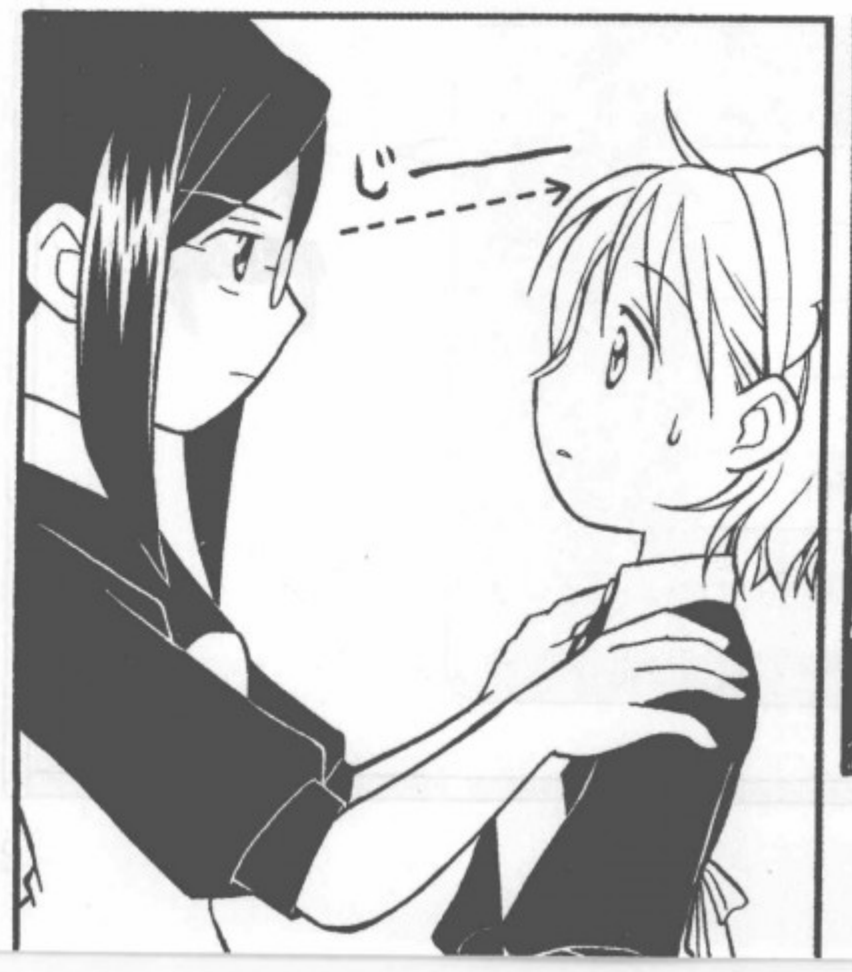
微分
啊哈！接下来我们学习微分！



对不起，我快要受不了啦……
哏！哏！



好，好的！
站起来！
没关系，不就是计算嘛！只要用心去做，就不会那么难啦！我给你好好讲讲，你要加油啊！！
ガタッ



じーー



156厘米！
是的，准确地说，是155.7厘米。
厉害！
当然！



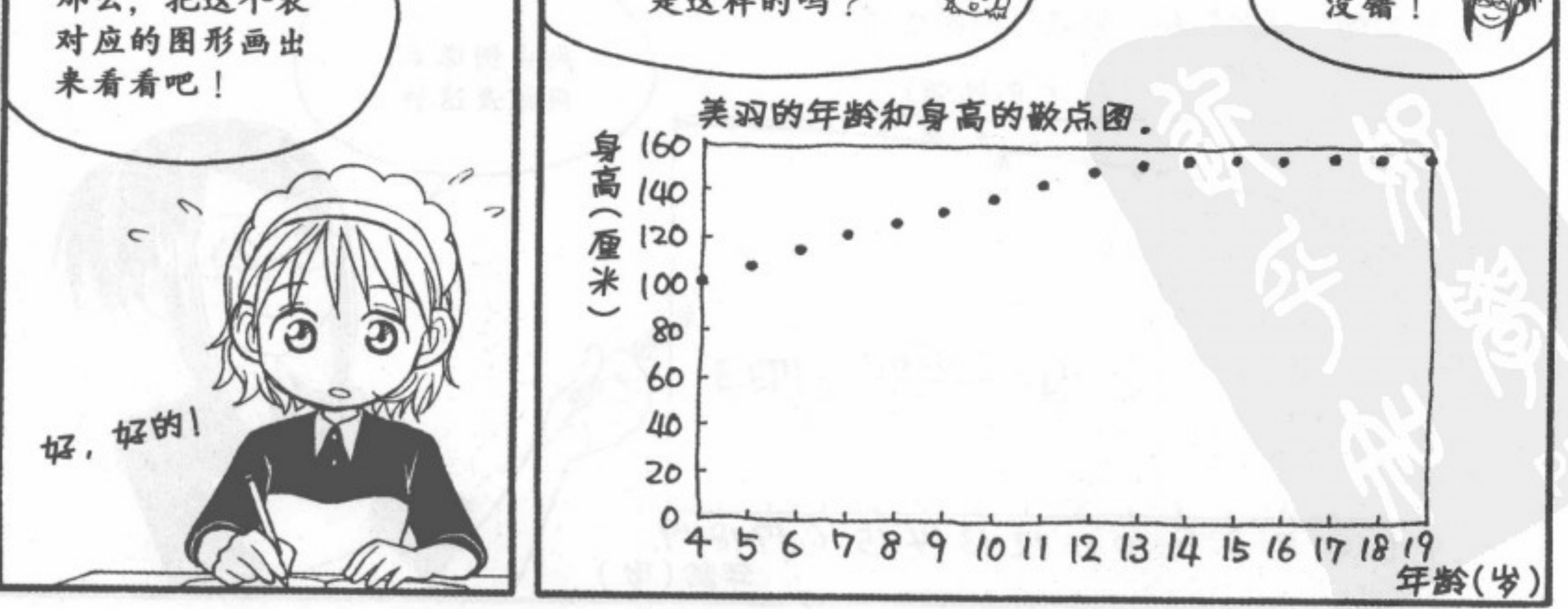
155.7厘米啊！！
155.7
カッ

美羽的"年龄"和"身高"

年龄(岁)	身高(厘米)
4	100.1
5	107.2
6	114.1
7	121.7
8	126.8
9	130.9
10	137.5
11	143.2
12	149.4
13	151.6
14	154.0
15	154.6
16	155.0
17	155.1
18	155.3
19	155.7

这是美羽从进入幼儿园开始到现在的身高数据。

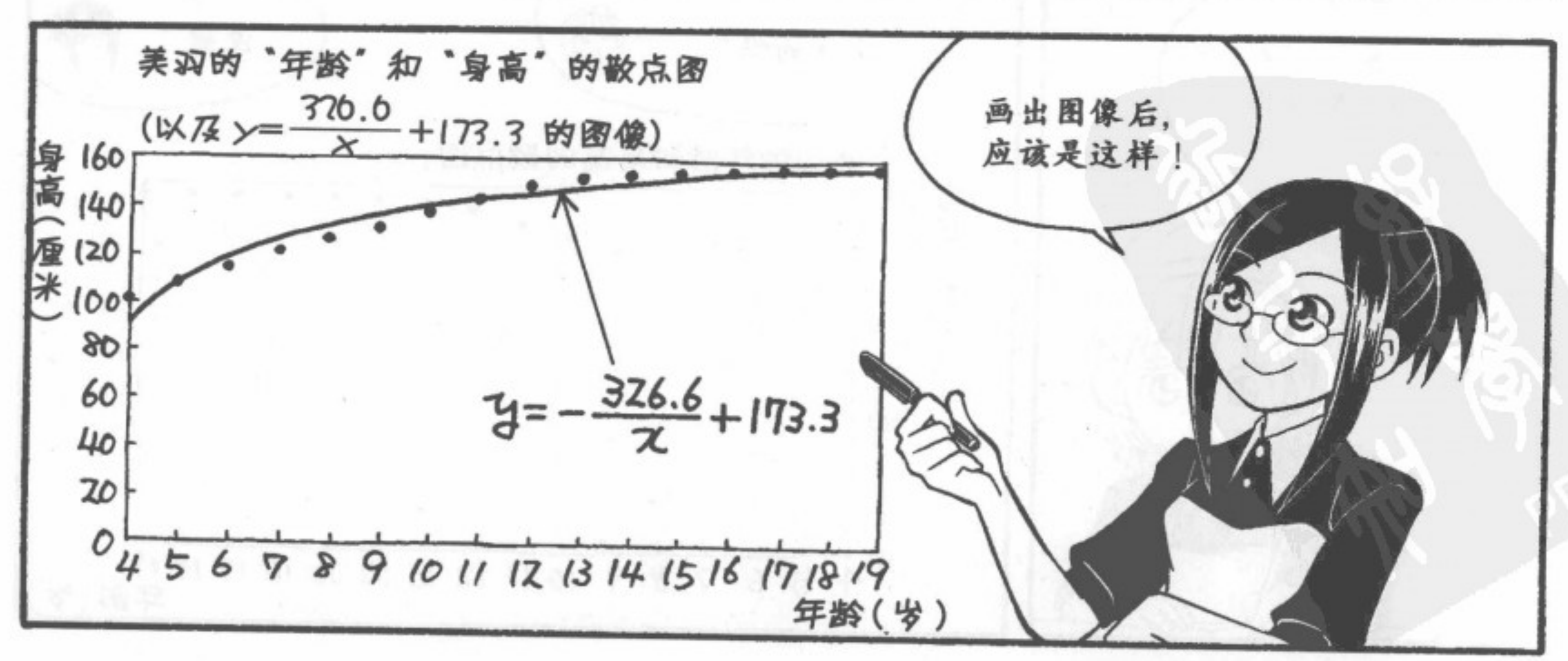





那么，让我们来看看6岁和7岁时的美羽吧！
114.1 cm
121.7 cm
6岁
7岁


美羽长个啦！
从6岁到7岁一年之间长了121.7－114.1=7.6cm。


然后，再回到刚才的问题，我们先抛开细节，大致地讲一下，美羽从4岁到19岁的“年龄”和“身高”之间……
$y=-\frac{326.6}{x}+173.3$
身高
年龄
……存在着这样的关系。

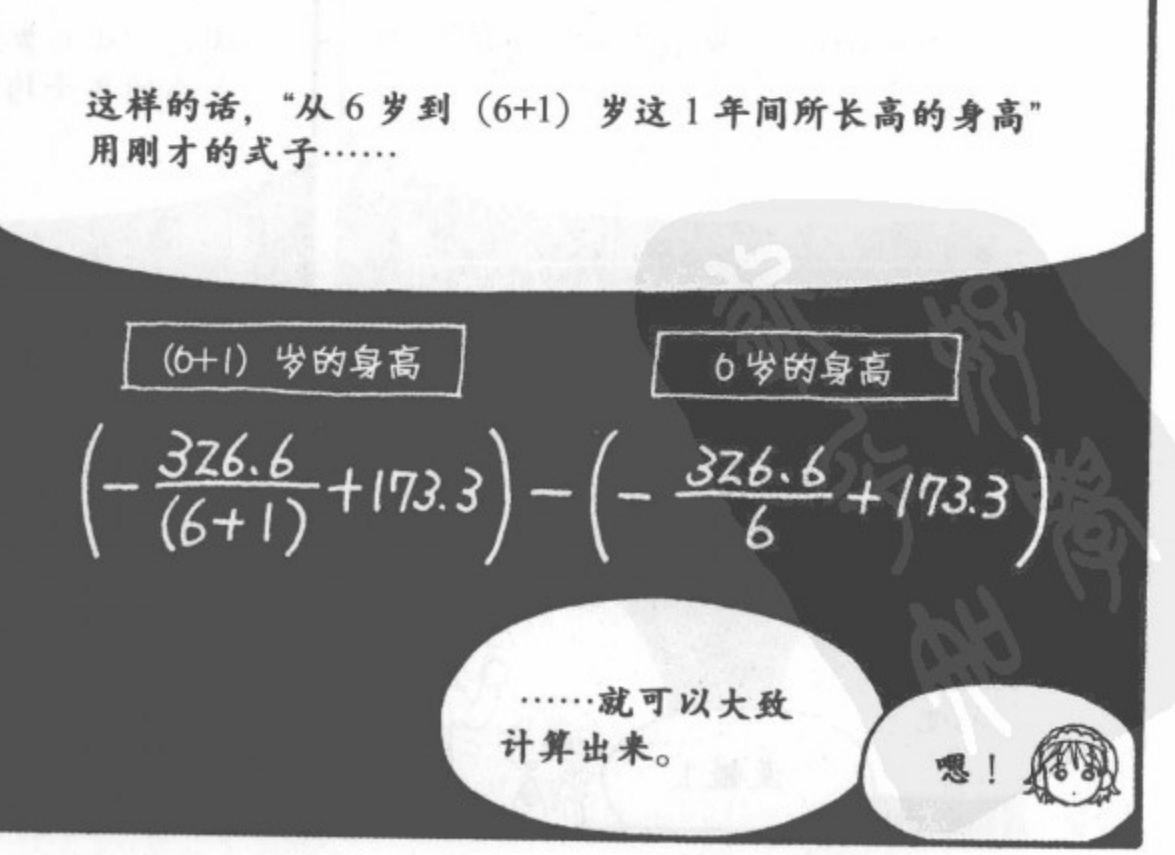

美羽的“年龄”和“身高”的散点图
(以及$y=-\frac{326.6}{x}+173.3$的图像)
身高(厘米)
160
140
120
100
80
60
40
20
0
4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
年龄(岁)
$y=-\frac{326.6}{x}+173.3$
画出图像后，应该是这样！

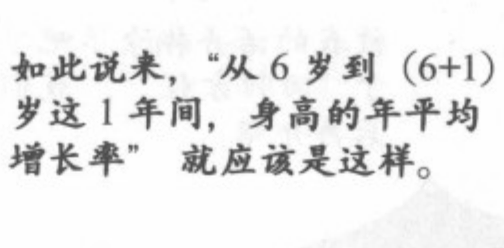

这，这个
$y=-\frac{326.6}{x}+173.3$
是什么啊？
有些复杂
……

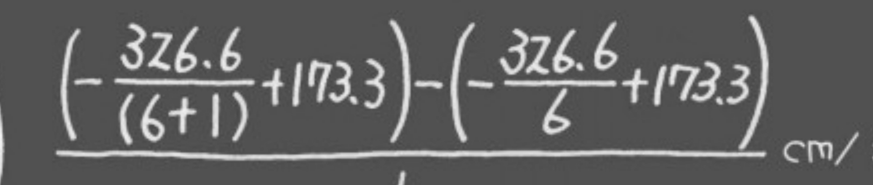

就是通过“回归分析”求解出来的“回归方程”！
被我的话弄糊涂了吧？关于“回归方程”，我们以后再介绍。


好吧！
嗯……
那就先把我的年龄和身高之间这个
$y=-\frac{326.6}{x}+173.3$
的关系弄清楚再说吧！
拜托了！


你看，我们是不是可以把“7岁”改写成“(6+1)岁”啊？
没错


这样的话，“从6岁到(6+1)岁这1年间所长高的身高”用刚才的式子……
(6+1)岁的身高
0岁的身高
$\left(-\frac{326.6}{(6+1)}+173.3\right)-\left(-\frac{326.6}{6}+173.3\right)$
……就可以大致计算出来。
嗯！

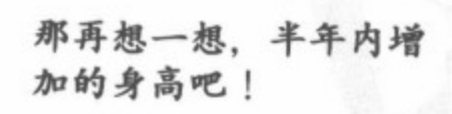


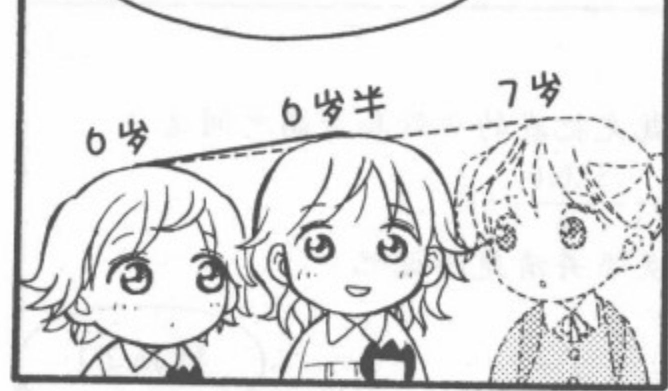


$$\frac{\left(-\frac{326.6}{(6+1)}+173.3\right)-\left(-\frac{326.6}{6}+173.3\right)}{1}\ \text{cm/年}$$











“从 6 岁到（6+0.5）岁这 0.5 年间所长高的身高”……

(6+0.5) 岁的身高　　6 岁的身高

$$\left(-\frac{326.6}{(6+0.5)}+173.3\right)-\left(-\frac{326.6}{6}+173.3\right)$$

大概是这样的吧！

美羽

没错！

那么，“从 6 岁到（6+0.5）岁这 0.5 年间，身高的年平均增长率”就是这样。

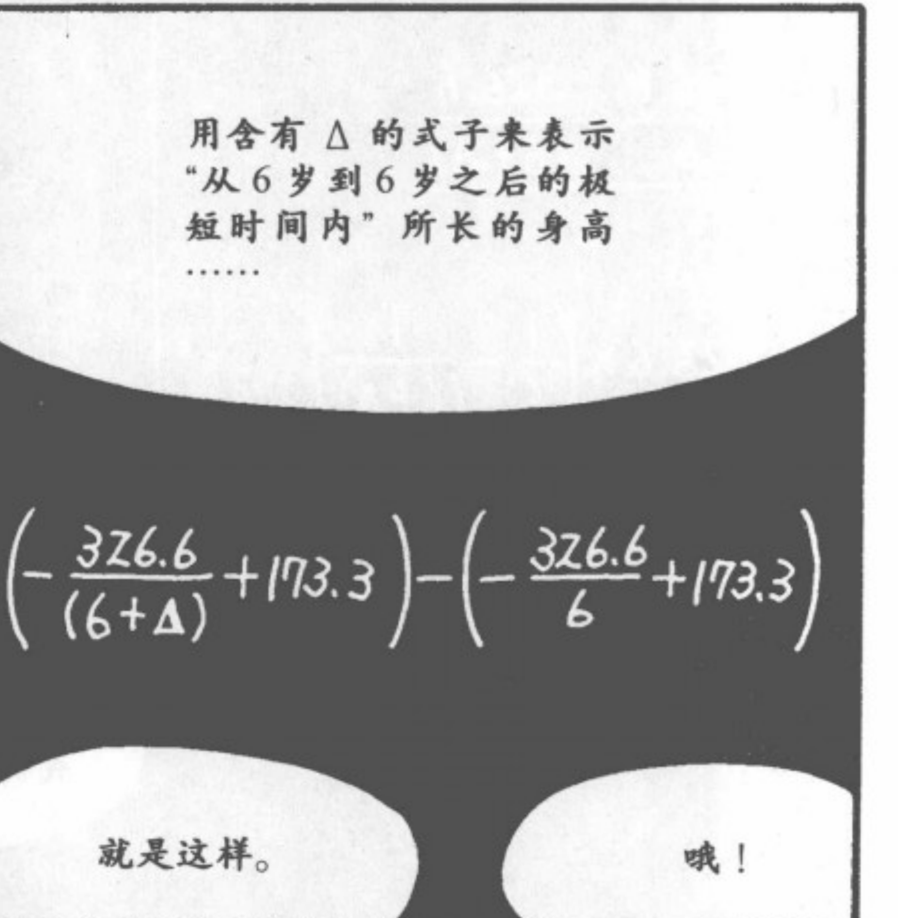
$$\frac{\left(-\frac{326.6}{(6+0.5)}+173.3\right)-\left(-\frac{326.6}{6}+173.3\right)}{0.5}\ \text{cm/年}$$

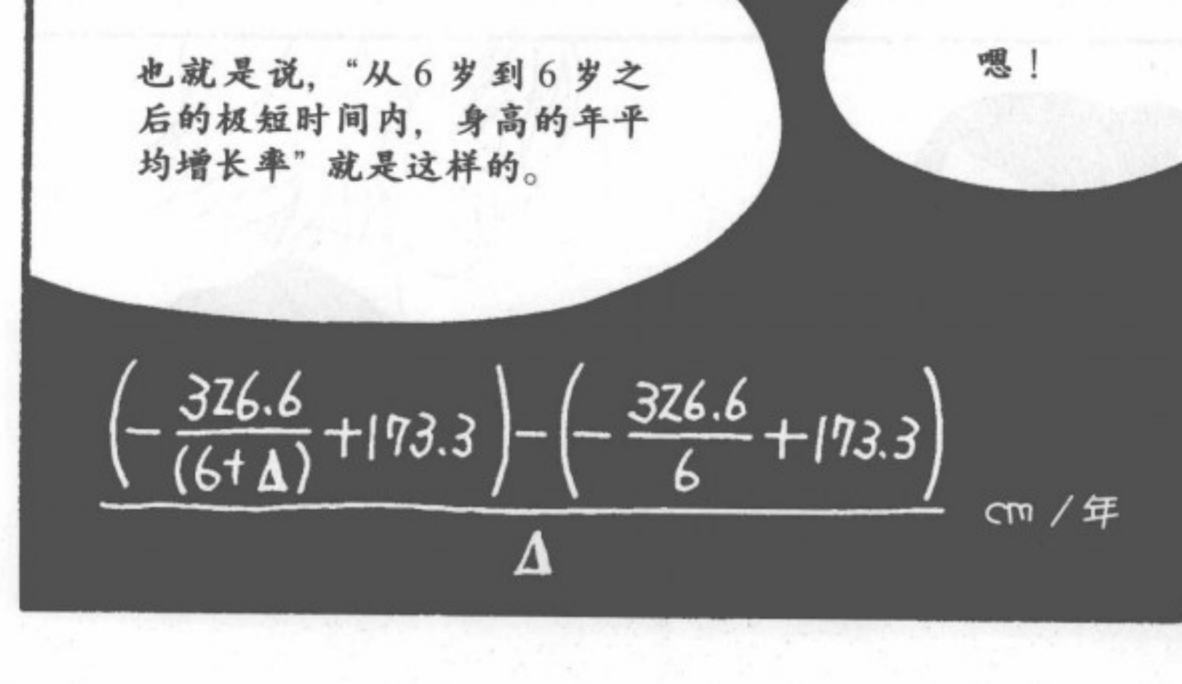
0.5 年间所长高的身高，再除以 0.5。


那么，最后……
考虑一下“极短时间”内，所长的身高。


“极短的时间”的话…
美羽
爸爸！


delte
Δ
没错！
在数学中，我们用 Δ (delta) 表示。


用含有 Δ 的式子来表示“从6岁到6岁之后的极短时间内”所长的身高……
$\left(-\frac{326.6}{(6+\Delta)}+173.3\right)-\left(-\frac{326.6}{6}+173.3\right)$
就是这样。
哦！


也就是说，“从6岁到6岁之后的极短时间内，身高的年平均增长率”就是这样的。
嗯！
$\frac{\left(-\frac{326.6}{(6+\Delta)}+173.3\right)-\left(-\frac{326.6}{6}+173.3\right)}{\Delta}$ cm/年


来吧，让我们一口气把这个式子整理出来！

$$\frac{\left(-\frac{326.6}{(6+\Delta)}+173.3\right)-\left(-\frac{326.6}{6}+173.3\right)}{\Delta}$$

$$=\frac{-\frac{326.6}{(6+\Delta)}+\frac{326.6}{6}}{\Delta}$$

$$=\frac{\frac{326.6}{6}-\frac{326.6}{(6+\Delta)}}{\Delta}$$

$$=\frac{326.6\times\left(\frac{1}{6}-\frac{1}{(6+\Delta)}\right)}{\Delta}$$

$$=\frac{326.6\times\frac{(6+\Delta)-6}{6(6+\Delta)}}{\Delta}$$

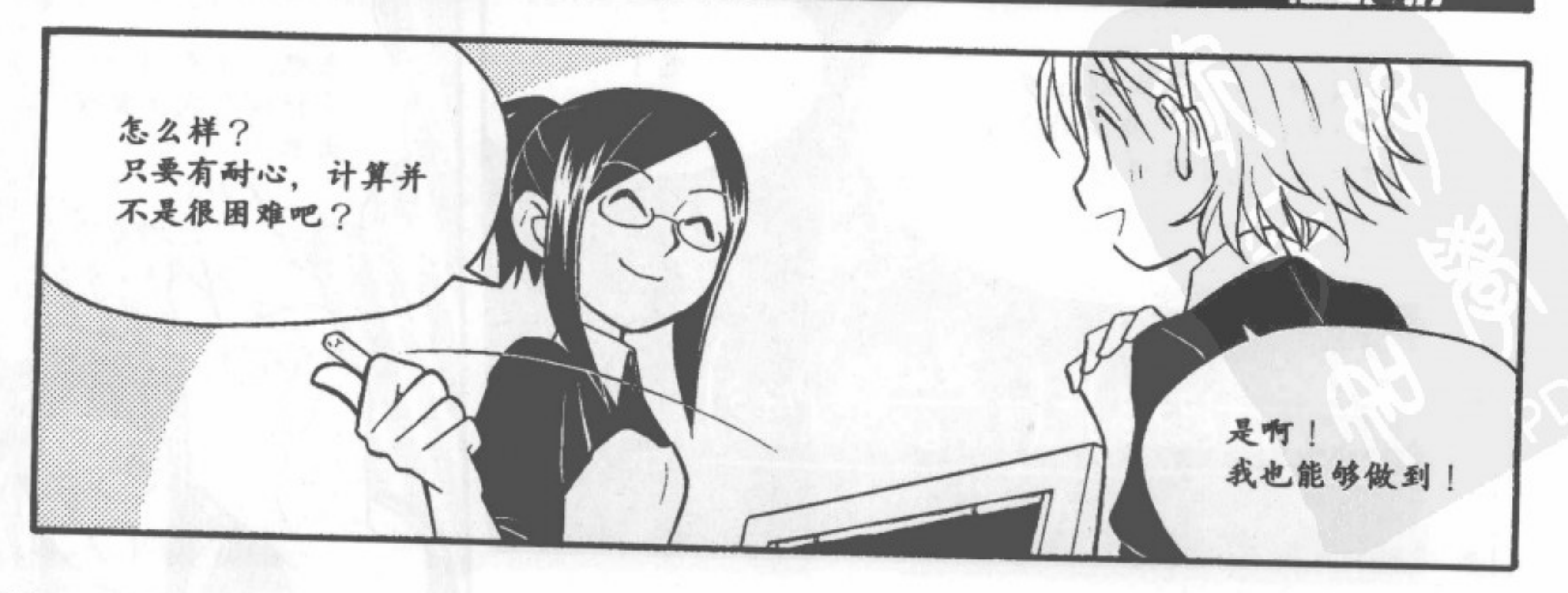
$$=\frac{326.6\times\frac{\Delta}{6(6+\Delta)}}{\Delta}$$

$$=326.6\times\frac{\Delta}{6(6+\Delta)}\times\frac{1}{\Delta}$$

$$=326.6\times\frac{1}{6(6+\Delta)}$$

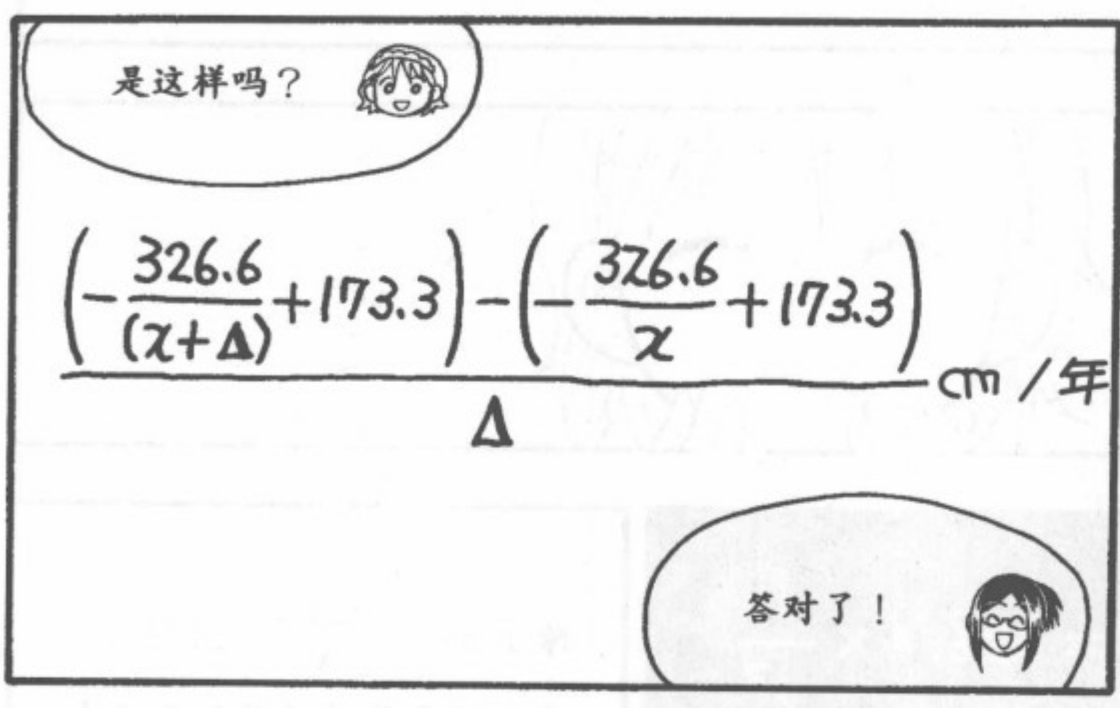
$$\fallingdotseq 326.6\times\frac{1}{6(6+0)}=326.6\times\frac{1}{6^2}\ \text{cm/年}$$

在“极短时间”的情况下，我们可以将 Δ 看作0。

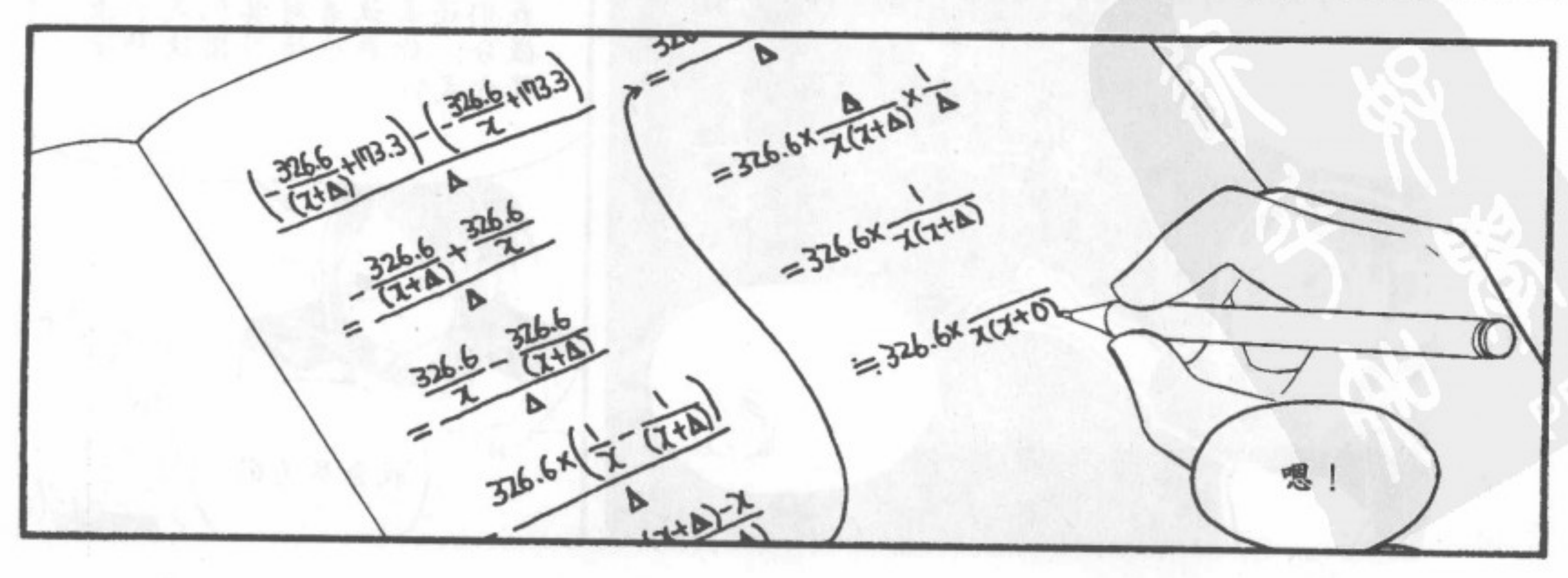



那就来接受一下
挑战吧！

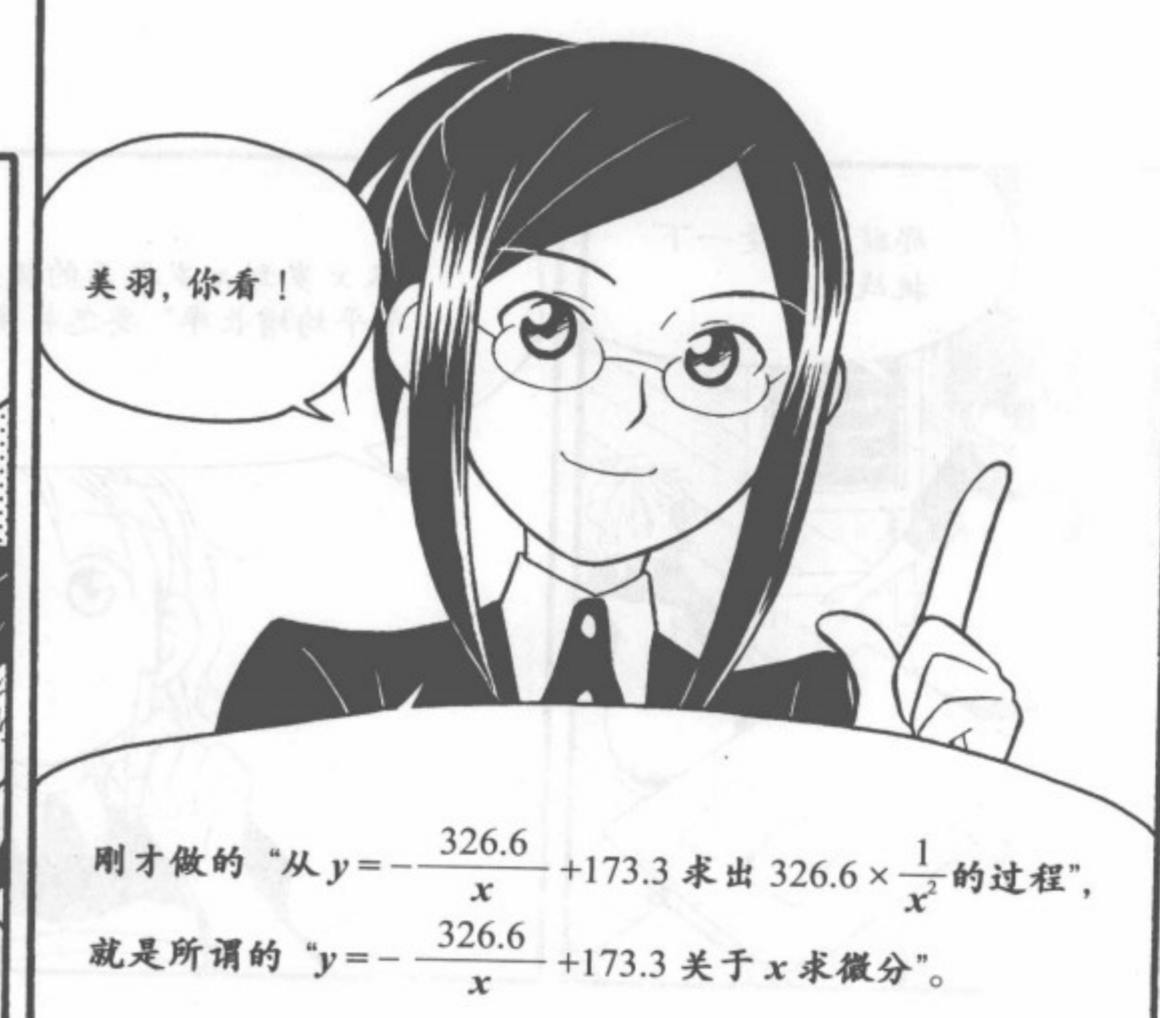

“从 x 岁到 x 岁之后的极短时间内，身高的
年平均增长率”要怎样用式子进行表示？
嗯！


是这样吗？
$$\frac{\left(-\frac{326.6}{(x+\Delta)}+173.3\right)-\left(-\frac{326.6}{x}+173.3\right)}{\Delta}$$ cm／年
答对了！

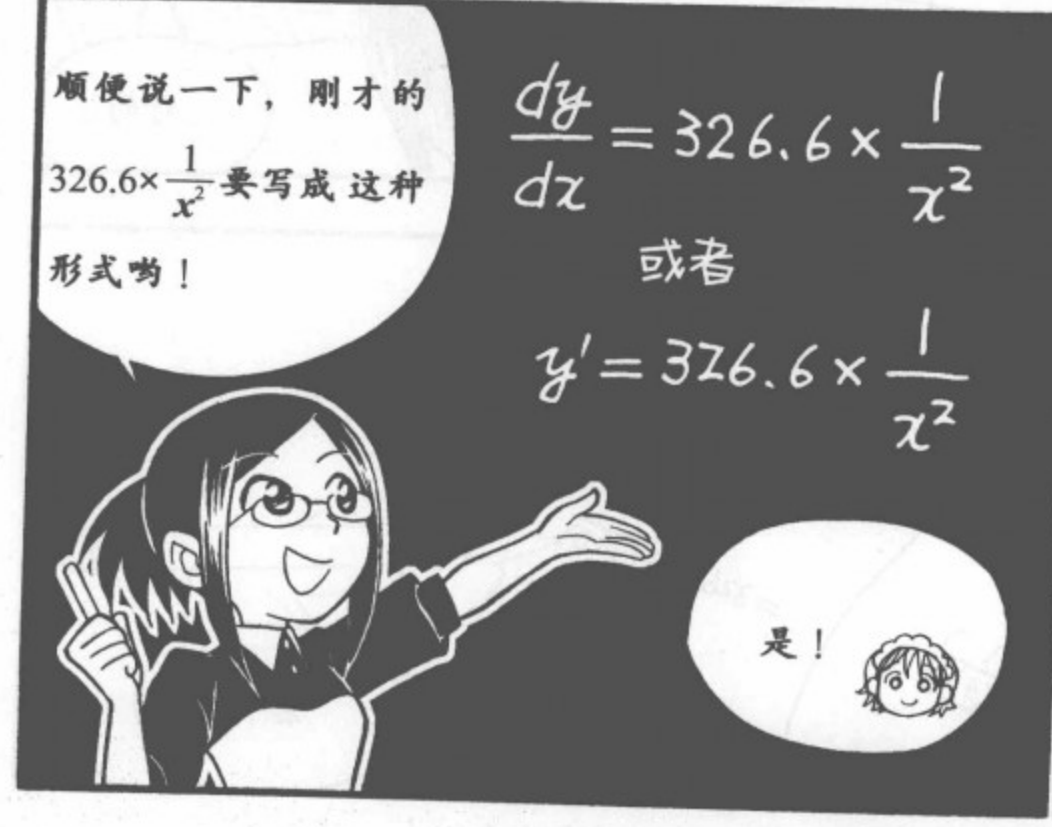

那就把它整理一
下，看看吧！
好的！


嗯！


最后变成 $326.6\times\frac{1}{x^2}$ 了！
回答正确！


美羽，你看！
刚才做的“从 $y=-\frac{326.6}{x}+173.3$ 求出 $326.6\times\frac{1}{x^2}$ 的过程”，
就是所谓的“$y=-\frac{326.6}{x}+173.3$ 关于 x 求微分”。


原来如此！


顺便说一下，刚才的 $326.6\times\frac{1}{x^2}$ 要写成这种形式哟！
$\frac{dy}{dx}=326.6\times\frac{1}{x^2}$
或者
$y'=326.6\times\frac{1}{x^2}$
是！


除了 $y=-\frac{326.6}{x}+173.3$，
我们还要试着对其他式子求微分，你要尽快习惯这种运算方式啊！
我会尽力的！



$y=x$, 关于 x 进行微分！



因为 $\dfrac{(x+\Delta)-x}{\Delta}+\dfrac{\Delta}{\Delta}=1$，所以 $\dfrac{dy}{dx}=1$。



$y=x^2$，关于 x 进行微分！



因为 $\dfrac{(x+\Delta)^2-x^2}{\Delta}=\dfrac{[(x+\Delta)+x][(x+\Delta)-x]}{\Delta}=\dfrac{(2x+\Delta)\times\Delta}{\Delta}=2x+\Delta$

$\approx 2x+0=2x$，所以 $\dfrac{dy}{dx}=2x$。



$y=\dfrac{1}{x}$，关于 x 进行微分！



因为 $\dfrac{\dfrac{1}{x+\Delta}-\dfrac{1}{x}}{\Delta}=\dfrac{\dfrac{x-(x+\Delta)}{(x+\Delta)x}}{\Delta}=\dfrac{\dfrac{-\Delta}{(x+\Delta)x}}{\Delta}=\dfrac{-\Delta}{(x+\Delta)x}\times\dfrac{1}{\Delta}$

$=\dfrac{-1}{(x+\Delta)x}=\dfrac{-1}{(x+0)x}$

$=\dfrac{-1}{x^2}=-x^{-2}$

所以，$\dfrac{dy}{dx}=-x^{-2}$。

$y=\dfrac{1}{x^2}$，关于 x 进行微分。





$$\frac{\frac{1}{(x+\Delta)^2}-\frac{1}{x^2}}{\Delta}$$

$$=\frac{\left(\frac{1}{x+\Delta}\right)^2-\left(\frac{1}{x}\right)^2}{\Delta}$$

$$=\frac{\left(\frac{1}{x+\Delta}+\frac{1}{x}\right)\left(\frac{1}{x+\Delta}-\frac{1}{x}\right)}{\Delta}$$

$$=\frac{\frac{x+(x+\Delta)}{(x+\Delta)x}\times\frac{x-(x+\Delta)}{(x+\Delta)x}}{\Delta}$$

$$=\frac{\frac{2x+\Delta}{(x+\Delta)x}\times\frac{-\Delta}{(x+\Delta)x}}{\Delta}$$

$$=\frac{2x+\Delta}{(x+\Delta)x}\times\frac{-\Delta}{(x+\Delta)x}\times\frac{1}{\Delta}$$

$$=\frac{-(2x+\Delta)}{[(x+\Delta)x]^2}$$

$$\approx\frac{-(2x+0)}{[(x+0)x]^2}$$

$$=\frac{-2x}{x^4}$$

$$=\frac{-2}{x^3}$$

$$=-2x^{-3}$$

因此，$\dfrac{dy}{dx}=-2x^{-3}$。

通过以上的例子，我们可以看出：

$y=x^n$ 关于 x 的微分，就是 $\dfrac{dy}{dx}=nx^{n-1}$。





$y=(5x-7)^2$，关于 x 进行微分。



$$\frac{[5(x+\Delta)-7]^2-(5x-7)^2}{\Delta}$$

$$=\frac{\{[5(x+\Delta)-7]+(5x-7)\}\{[5(x+\Delta)-7]-(5x-7)\}}{\Delta}$$

$$=\frac{[2(5x-7)+5\Delta]\times 5\Delta}{\Delta}$$

$$=[2(5x-7)+5\Delta]\times 5$$

$$\approx [2(5x-7)+5\times 0]\times 5$$

$$=2(5x-7)\times 5$$

因此，$\frac{dy}{dx}=2(5x-7)\times 5$。

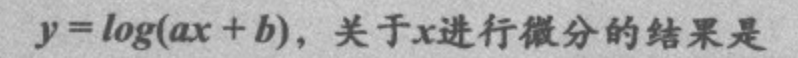


类似 $y=(ax+b)^n$ 这种形式，关于 x 进行微分的结果就是 $\frac{dy}{dx}=n(ax+b)^{n-1}\times a$。

由于计算繁杂，这里就不一一讲解了，不过类似这样的还有…

$y=e^x$，关于x进行微分的结果是$\dfrac{dy}{dx}=e^x$。

$y=log\,x$，关于x进行微分的结果是$\dfrac{dy}{dx}=\dfrac{1}{x}$。



$y=log(ax+b)$，关于x进行微分的结果是

$\dfrac{dy}{dx}=\dfrac{1}{ax+b}\times a$。

$y=log(1+ea^{x+b})$，关于x进行微分的结果是

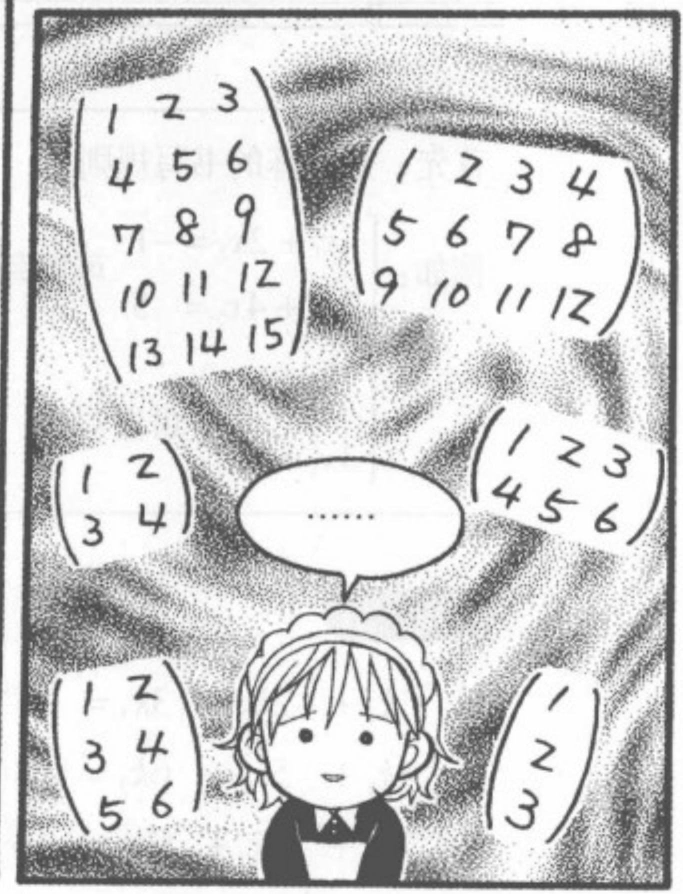
$\dfrac{dy}{dx}=\dfrac{1}{1+e^{ax+b}}\times ae^{ax+b}$。



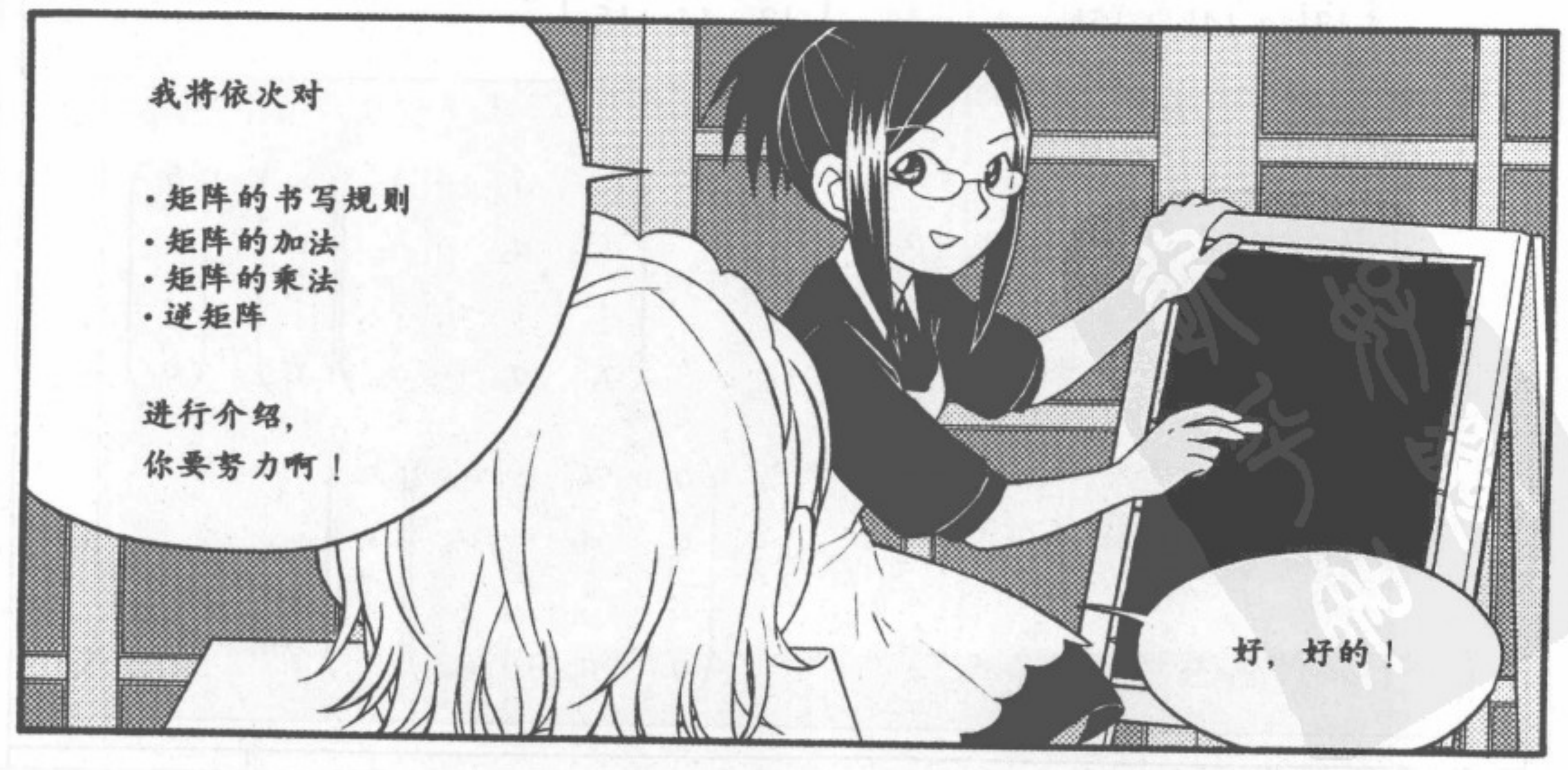


6. 矩　阵


矩阵
好吧！
那么，最后
我们来学习
矩阵！！


……


看起来像用数字盖的
大楼啊……
矩阵也不会吗……


我将依次对
·矩阵的书写规则
·矩阵的加法
·矩阵的乘法
·逆矩阵
进行介绍，
你要努力啊！
好，好的！

首先，是矩阵的书写规则。

例如：$\begin{cases} x_1 + 2x_2 = -1 \\ 3x_1 + 4x_2 = \ \ 5 \end{cases}$ 可以写作：$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1 \\ 5 \end{pmatrix}$

$\begin{cases} x_1 + 2x_2 \\ 3x_1 + 4x_2 \end{cases}$ 可以写作：$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$



例如：

$$\begin{cases} k_1 + \ 2k_2 + \ 3k_3 = -3 \\ 4k_1 + \ 5k_2 + \ 6k_3 = \ \ 8 \\ 7k_1 + \ 8k_2 + \ 9k_3 = \ \ 6 \\ 10k_1 + 11k_2 + 12k_3 = \ \ 2 \\ 13k_1 + 14k_2 + 15k_3 = \ \ 7 \end{cases} \text{可以写作：} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix}\begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix} = \begin{pmatrix} -3 \\ 8 \\ 6 \\ 2 \\ 7 \end{pmatrix}$$

$$\begin{cases} k_1 + \ 2k_2 + \ 3k_3 \\ 4k_1 + \ 5k_2 + \ 6k_3 \\ 7k_1 + \ 8k_2 + \ 9k_3 \\ 10k_1 + 11k_2 + 12k_3 \\ 13k_1 + 14k_2 + 15k_3 \end{cases} \text{可以写作：} \begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix}\begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix}$$

归纳如下：

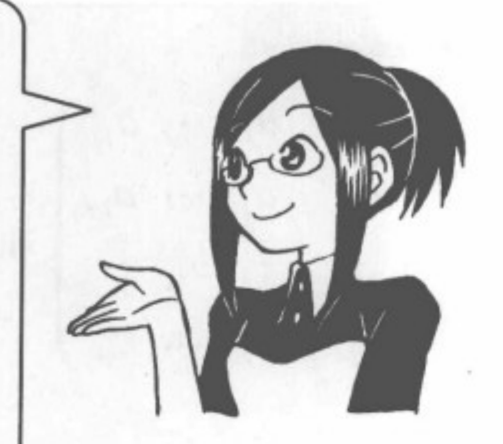
- $\begin{cases} a_{11}x_1 + a_{12}x_2 + \ \cdots + a_{1q}x_q = b_1 \\ a_{21}x_2 + a_{22}x_2 + \ \cdots + a_{2q}x_q = b_2 \\ \cdots\cdots \\ a_{p1}x_1 + a_{p2}x_2 + \ \cdots + a_{pq}x_q = b_p \end{cases}$ 可以写作：$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$

- $\begin{cases} a_{11}x_1 + a_{12}x_2 + \ \cdots + a_{1q}x_q \\ a_{21}x_1 + a_{22}x_2 + \ \cdots + a_{2q}x_q \\ \cdots\cdots \\ a_{p1}x_1 + a_{p2}x_2 + \ \cdots + a_{pq}x_q \end{cases}$ 可以写作：$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_q \end{pmatrix}$

其次,我们来讲一下矩阵的加法:

例如: $\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}$ 和 $\begin{pmatrix}4 & 5\\-2 & 4\end{pmatrix}$ 相加, $\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}+\begin{pmatrix}4 & 5\\-2 & 4\end{pmatrix}$

的计算可按如下方式进行: $\begin{pmatrix}1+4 & 2+5\\3+(-2) & 4+4\end{pmatrix}$



例 1

$$\begin{pmatrix}5 & 1\\6 & -9\end{pmatrix}+\begin{pmatrix}-1 & 3\\-3 & 10\end{pmatrix}$$

可按如下方式计算: $\begin{pmatrix}5 & 1\\6 & -9\end{pmatrix}+\begin{pmatrix}-1 & 3\\-3 & 10\end{pmatrix}=\begin{pmatrix}5+(-1) & 1+3\\6+(-3) & (-9)+10\end{pmatrix}=\begin{pmatrix}4 & 4\\3 & 1\end{pmatrix}$

例 2

$$\begin{pmatrix}1 & 2 & 3\\4 & 5 & 6\\7 & 8 & 9\\10 & 11 & 12\\13 & 14 & 15\end{pmatrix}+\begin{pmatrix}7 & 2 & 3\\-1 & 7 & -4\\-7 & -3 & 10\\8 & 2 & -1\\7 & 1 & -9\end{pmatrix}$$

可按如下方式计算:

$$\begin{pmatrix}1 & 2 & 3\\4 & 5 & 6\\7 & 8 & 9\\10 & 11 & 12\\13 & 14 & 15\end{pmatrix}+\begin{pmatrix}7 & 2 & 3\\-1 & 7 & -4\\-7 & -3 & 10\\8 & 2 & -1\\7 & 1 & -9\end{pmatrix}=\begin{pmatrix}1+7 & 2+2 & 3+3\\4+(-1) & 5+7 & 6+(-4)\\7+(-7) & 8+(-3) & 9+10\\10+8 & 11+2 & 12+(-1)\\13+7 & 14+1 & 15+(-9)\end{pmatrix}=\begin{pmatrix}8 & 4 & 6\\3 & 12 & 2\\0 & 5 & 19\\18 & 13 & 11\\20 & 15 & 6\end{pmatrix}$$

归纳如下：

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix} \text{和} \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1q} \\ b_{21} & b_{22} & \cdots & b_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pq} \end{pmatrix} \text{相加}$$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1q} \\ b_{21} & b_{22} & \cdots & b_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pq} \end{pmatrix} \text{可以写作}$$

$$\begin{pmatrix} a_{11}+b_{11} & a_{12}+b_{12} & \cdots & a_{1q}+b_{1q} \\ a_{21}+b_{21} & a_{22}+b_{22} & \cdots & a_{2q}+b_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}+b_{p1} & a_{p2}+b_{p2} & \cdots & a_{pq}+b_{pq} \end{pmatrix}$$

接下来，我们讲讲矩阵的乘法。

$\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}\begin{pmatrix}x_1 & y_1\\x_2 & y_2\end{pmatrix}$ 虽说是“乘法”，其实就是将

$\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}\begin{pmatrix}x_1\\x_2\end{pmatrix}$ 和 $\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}\begin{pmatrix}y_1\\y_2\end{pmatrix}$ 单独运算，也就是表示同时计算

$\begin{cases}x_1+2x_2\\3x_1+4x_2\end{cases}$ 和 $\begin{cases}y_1+2y_2\\3y_1+4y_2\end{cases}$



例 1

$\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}\begin{pmatrix}4 & 5\\-2 & 4\end{pmatrix}$ 可进行如下计算：

- $\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}\begin{pmatrix}4\\-2\end{pmatrix}=\begin{pmatrix}1\times4+2\times(-2)\\3\times4+4\times(-2)\end{pmatrix}=\begin{pmatrix}0\\4\end{pmatrix}$

- $\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}\begin{pmatrix}5\\4\end{pmatrix}=\begin{pmatrix}1\times5+2\times4\\3\times5+4\times4\end{pmatrix}=\begin{pmatrix}13\\31\end{pmatrix}$

因此，可得 $\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}\begin{pmatrix}4 & 5\\-2 & 4\end{pmatrix}=\begin{pmatrix}0 & 13\\4 & 31\end{pmatrix}$

例 2

$\begin{pmatrix}1 & 2 & 3\\4 & 5 & 6\\7 & 8 & 9\\10 & 11 & 12\\13 & 14 & 15\end{pmatrix}\begin{pmatrix}k_1 & l_1 & m_1 & n_1\\k_2 & l_2 & m_2 & n_2\\k_3 & l_3 & m_3 & n_3\end{pmatrix}$ 的计算：

· $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix}\begin{pmatrix} k_1 \\ k_2 \\ k_3 \end{pmatrix}=\begin{pmatrix} k_1+2k_2+3k_3 \\ 4k_1+5k_2+6k_3 \\ 7k_1+8k_2+9k_3 \\ 10k_1+11k_2+12k_3 \\ 13k_1+14k_2+15k_3 \end{pmatrix}$

· $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix}\begin{pmatrix} l_1 \\ l_2 \\ l_3 \end{pmatrix}=\begin{pmatrix} l_1+2l_2+3l_3 \\ 4l_1+5l_2+6l_3 \\ 7l_1+8l_2+9l_3 \\ 10l_1+11l_2+12l_3 \\ 13l_1+14l_2+15l_3 \end{pmatrix}$

· $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix}\begin{pmatrix} m_1 \\ m_2 \\ m_3 \end{pmatrix}=\begin{pmatrix} m_1+2m_2+3m_3 \\ 4m_1+5m_2+6m_3 \\ 7m_1+8m_2+9m_3 \\ 10m_1+11m_2+12m_3 \\ 13m_1+14m_2+15m_3 \end{pmatrix}$

· $\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \\ 10 & 11 & 12 \\ 13 & 14 & 15 \end{pmatrix}\begin{pmatrix} n_1 \\ n_2 \\ n_3 \end{pmatrix}=\begin{pmatrix} n_1+2n_2+3n_3 \\ 4n_1+5n_2+6n_3 \\ 7n_1+8n_2+9n_3 \\ 10n_1+11n_2+12n_3 \\ 13n_1+14n_2+15n_3 \end{pmatrix}$

因此，可得：

$$\begin{pmatrix} k_1+2k_2+3k_3 & l_1+2l_2+3l_3 & m_1+2m_2+3m_3 & n_1+2n_2+3n_3 \\ 4k_1+5k_2+6k_3 & 4l_1+5l_2+6l_3 & 4m_1+5m_2+6m_3 & 4n_1+5n_2+6n_3 \\ 7k_1+8k_2+9k_3 & 7l_1+8l_2+9l_3 & 7m_1+8m_2+9m_3 & 7n_1+8n_2+9n_3 \\ 10k_1+11k_2+12k_3 & 10l_1+11l_2+12l_3 & 10m_1+11m_2+12m_3 & 10n_1+11n_2+12n_3 \\ 13k_1+14k_2+15k_3 & 13l_1+14l_2+15l_3 & 13m_1+14m_2+15m_3 & 13n_1+14n_2+15n_3 \end{pmatrix}$$

归纳如下：

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix} \text{和} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1r} \\ x_{21} & x_{22} & \cdots & x_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ x_{q1} & x_{q2} & \cdots & x_{qr} \end{pmatrix} \text{的乘法计算，}$$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1r} \\ x_{21} & x_{22} & \cdots & x_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ x_{q1} & x_{q2} & \cdots & x_{qr} \end{pmatrix}$$

也就是

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix} \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{q1} \end{pmatrix} \text{和} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix} \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{q2} \end{pmatrix} \text{和……和}$$

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1q} \\ a_{21} & a_{22} & \cdots & a_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pq} \end{pmatrix} \begin{pmatrix} x_{1r} \\ x_{2r} \\ \vdots \\ x_{qr} \end{pmatrix} \text{分别运算，}$$

$$\text{也就是表示} \begin{cases} a_{11}x_{11} + a_{12}x_{21} + \cdots + a_{1q}x_{q1} \\ a_{21}x_{11} + a_{22}x_{21} + \cdots + a_{2q}x_{q1} \\ \cdots\cdots \\ a_{p1}x_{11} + a_{p2}x_{21} + \cdots + a_{pq}x_{q1} \end{cases} \text{和} \begin{cases} a_{11}x_{12} + a_{12}x_{22} + \quad + a_{1q}x_{q2} \\ a_{21}x_{12} + a_{22}x_{22} + \quad + a_{2q}x_{q2} \\ \cdots\cdots \\ a_{p1}x_{12} + a_{p2}x_{22} + \quad + a_{pq}x_{q2} \end{cases} \text{和……和}$$

$$\begin{cases} a_{11}x_{1r} + a_{12}x_{2r} + \cdots + a_{1q}x_{qr} \\ a_{21}x_{1r} + a_{22}x_{2r} + \cdots + a_{2q}x_{qr} \\ \cdots\cdots \\ a_{p1}x_{1r} + a_{p2}x_{2r} + \cdots + a_{pq}x_{qr} \end{cases} \text{同时进行计算。}$$

最后，我再对逆矩阵做一些讲解：

例如，$\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}$ 的逆矩阵 $\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}^{-1}$，就是和 $\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}$ 相乘后得到 $\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$ 的那个矩阵。



例

$$\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}\begin{pmatrix}-2 & 1\\1.5 & -0.5\end{pmatrix}=\begin{pmatrix}1\times(-2)+2\times1.5 & 1\times1+2\times(-0.5)\\3\times(-2)+4\times1.5 & 3\times1+4\times(-0.5)\end{pmatrix}=\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}$$

因此 $\begin{pmatrix}-2 & 1\\1.5 & -0.5\end{pmatrix}=\begin{pmatrix}1 & 2\\3 & 4\end{pmatrix}^{-1}$

归纳如下：

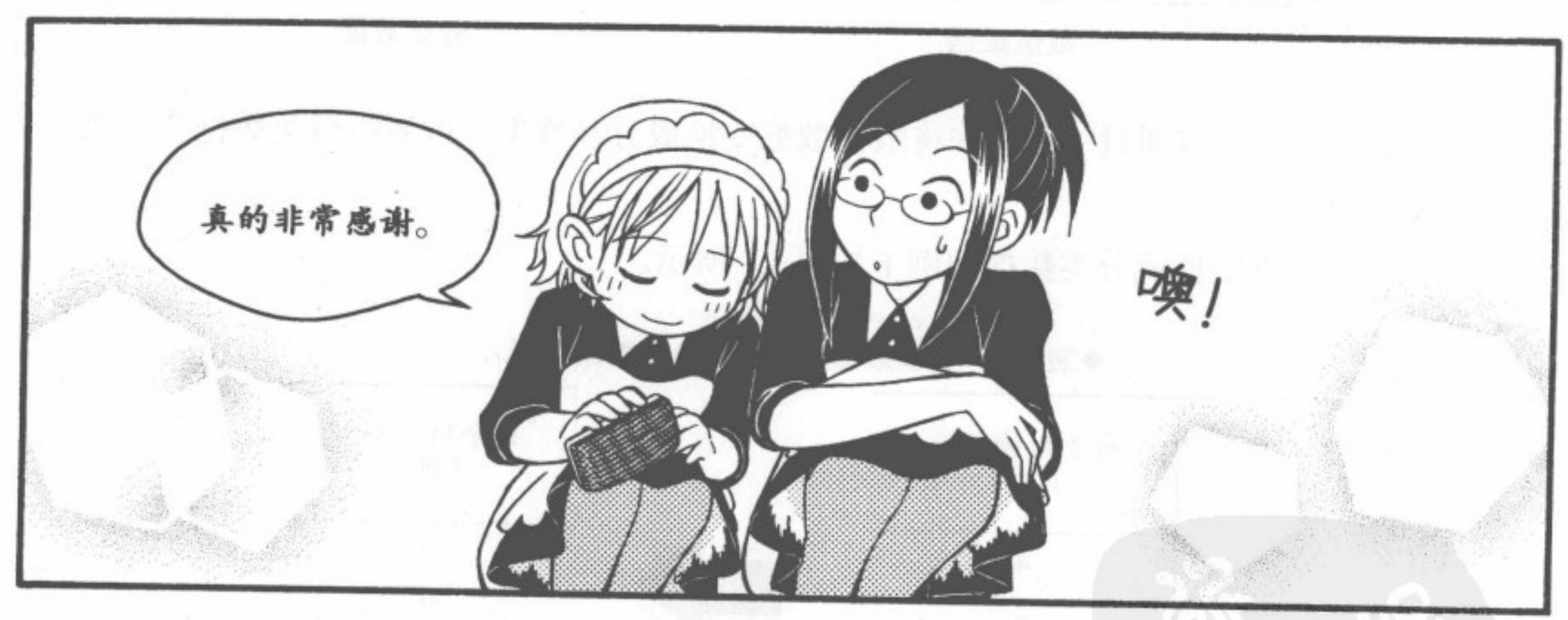
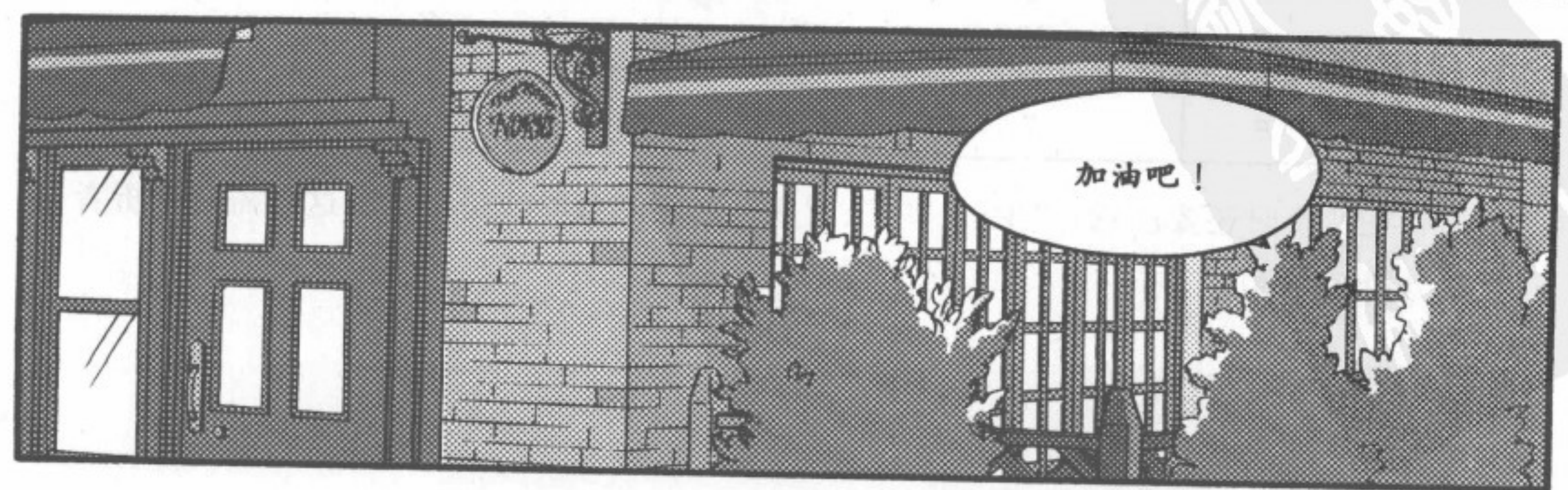
$$\begin{pmatrix}a_{11} & a_{12} & \cdots & a_{1p}\\a_{21} & a_{22} & \cdots & a_{2p}\\\vdots & \vdots & \ddots & \vdots\\a_{p1} & a_{p2} & \cdots & a_{pp}\end{pmatrix}$$ 的逆矩阵 $$\begin{pmatrix}a_{11} & a_{12} & \cdots & a_{1p}\\a_{21} & a_{22} & \cdots & a_{2p}\\\vdots & \vdots & \ddots & \vdots\\a_{p1} & a_{p2} & \cdots & a_{pp}\end{pmatrix}^{-1}$$ 就是与

$$\begin{pmatrix}a_{11} & a_{12} & \cdots & a_{1p}\\a_{21} & a_{22} & \cdots & a_{2p}\\\vdots & \vdots & \ddots & \vdots\\a_{p1} & a_{p2} & \cdots & a_{pp}\end{pmatrix}$$ 相乘后得 $$\begin{pmatrix}1 & 0 & \cdots & 0\\0 & 1 & \cdots & 0\\\vdots & \vdots & \ddots & \vdots\\0 & 0 & \cdots & 1\end{pmatrix}$$ 的那个矩阵。




要好好复习啊！
下次就要讲你所期待的回归分析了！
ガタッ

啊，我！

理纱前辈……

真的非常感谢。
噢！

加油吧！

✿ 7. 数值数据和分类数据 ✿

数值大致可以分为“可测量”数据和“不可测量”数据两种。其中，“可测量”数据被称为数值数据（numerical data），“不可测量”数据被称为分类数据（category data 或 categorial data）。

数值数据和分类数据在记录上的具体差别如表 1.1 所示。

◆表 1.1　数值数据和分类数据的具体实例

	每个月读书量（册）	年龄（岁）	主要的读书场所	性别
A	4	20	火车	女
B	2	19	家	男
C	10	18	咖啡厅	男
D	14	22	图书馆	女
⋮	⋮	⋮	⋮	⋮
	数值数据		分类数据	

分析者也可以通过一些方法将数值数据变换成分类数据，或者将分类数据变换成数值数据。

数值数据变换成分类数据的例子如表 1.2 所示。

◆表 1.2　数值数据变换成为分类数据举例

	每个月读书量（册）
A	4
B	2
C	10
D	14
E	7

➡

每个月读书量（册）
少
少
多
多
中

在变换时还要注意，“少”、“中”、“多”之间的分界线是多少，这是需要分析者做出相应判断的。

分类数据变换成为数值数据的例子如表 1.3 所示。

◆表 1.3 分类数据变换成为数值数据举例

	最喜欢的季节
A	春
B	夏
C	秋
D	冬

➡

春	夏	秋	冬
1	0	0	0
0	1	0	0
0	0	1	0
0	0	0	1

这里再对分类数据变换成为数值数据的例子稍作说明。对于上表中的变换，在实际操作中我们一般采用下表的形式。

◆表 1.4 分类数据变换成为数值数据举例（3 列）

	最喜欢的季节
A	春
B	夏
C	秋
D	冬

➡

春	夏	秋
1	0	0
0	1	0
0	0	1
0	0	0

同样地,例如我们在变换“星期几”时会采用 6 列,在变换“月份”时采用 11 列,在变换“性别”时采用 1 列。在表 1.4 中,虽然我们省略的是表 1.3 中“冬”的那一列,但是,如果我们不省略“冬”而省略“夏”的那一列也是正确的。以此类推,省略“春”或者“秋”都没有关系。

为什么特意省略了一列呢？是由于以下原因：

• 如果没有省略其中某一列，而直接使用像表 1.3 那样的数据进行重回归分析的话，在数学上是没有办法求解的。

• 虽然省略了一列，但是表达的意思是一样的（例如表 1.4 中的“冬”可以通过“0-0-0”进行表示）。再进一步讲，那一列的存在本身也是没有意义的，即便省略了也是合理的。

✿ 8. 离差平方和、方差、标准差 ✿

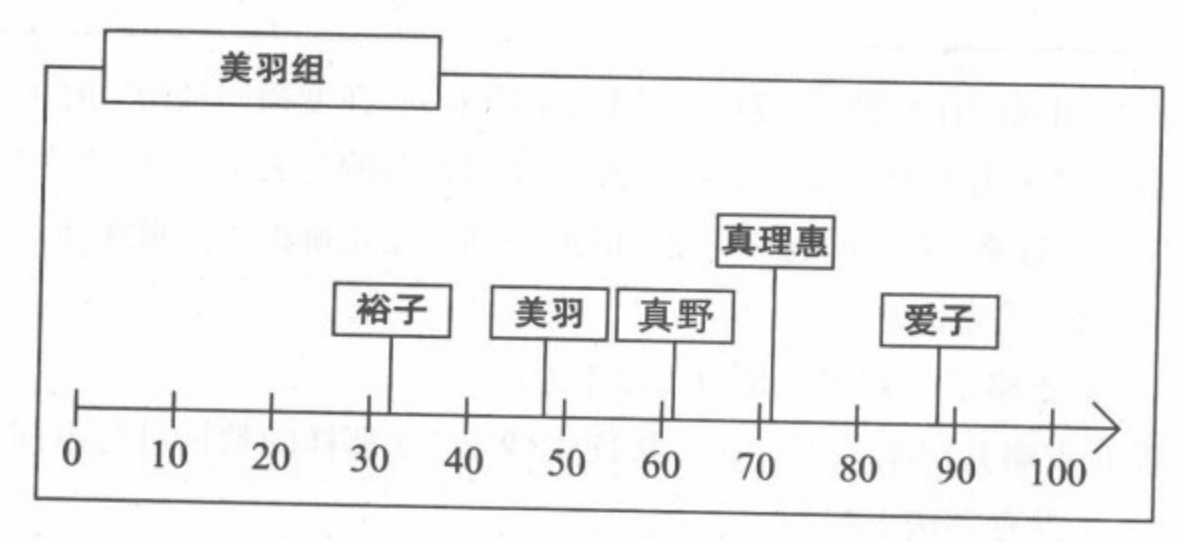
美羽和理纱与打工的同伴一起去唱卡拉 OK。她们每 5 人一组，分成两组依据演唱得分进行比赛。比赛结果如表 1.5。

◆表 1.5 卡拉 OK 的评分结果

	美羽组（得分）
美羽	48
裕子	32
爱子	88
真野	61
真理惠	71
平均	60

	理纱组（得分）
理纱	67
明日香	55
奈奈	61
雪儿	63
丽香	54
平均	60

将上表画成图便可得到下图。



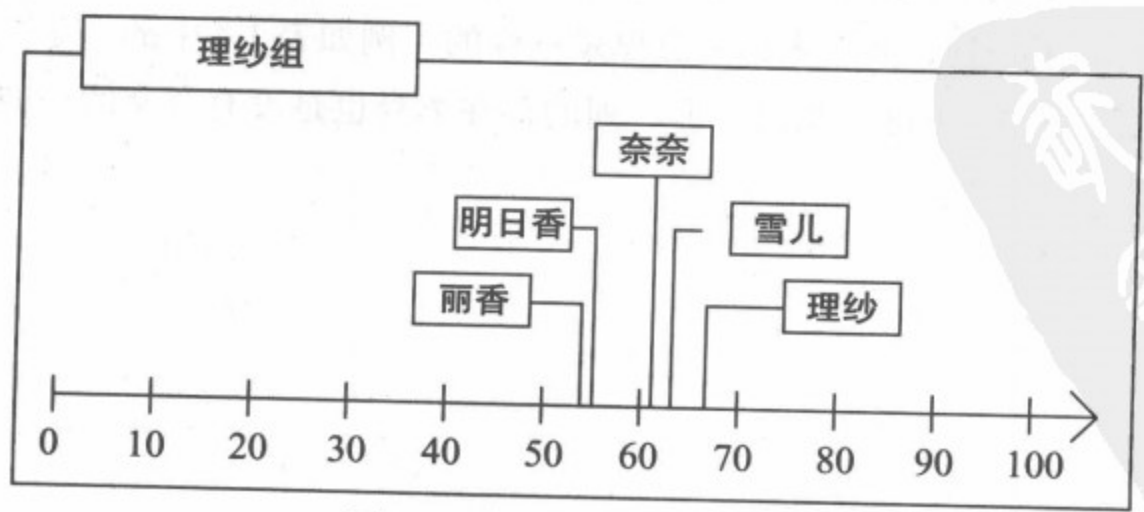


图 1.1 卡拉 OK 的评分结果

虽然美羽组和理纱组的平均得分同为 60 分，但具体的情况却差别很大。美羽组这方，每个人的得分是不是分布得很不均匀呢？也就是说数据的“分散程度”比较大。

人们通常采用离差平方和、(总体)方差和(总体)标准差作为表征数据的“离散程度”的指标。这些指标都具有如下性质：

- 最小值为 0
- 数据的“离散程度”越大，它们的值也就越大

离差平方和，常常会出现在以回归分析为代表的多种分析方法的计算过程中。

$$\text{离差平方和} = (\text{每个数据} - \text{平均值})^2\text{相加之和}$$

通过上述计算便可求解出离差平方和的值。然而数据的个数越多，它的值也就会变得越大，这也成为它的一个致命的缺点，所以在实际操作中我们很少使用它作为表征“离散程度”的指标。

(总体)方差，解决了离差平方和的缺点。可以通过如下计算求得它的值。

$$(\text{总体})\text{方差} = \frac{\text{离差平方和}}{\text{数据的个数}}$$

(总体)标准差，从本质上讲与(总体)方差是相同的。可以通过如下计算求得它的值。

$$(\text{总体})\text{标准差} = \sqrt{(\text{总体})\text{方差}}$$

让我们来求一下美羽组和理纱组的离差平方和、(总体)方差和(总体)标准差吧！

◆表 1.6　美羽组和理纱组的离差平方和、(总体)方差和(总体)标准差

	美羽组	理纱组
离差平方和	$(48-60)^2+(32-60)^2+(88-60)^2+(61-60)^2+(71-60)^2$ $=(-12)^2+(-28)^2+28^2+1^2+11^2$ $=1834$	$(67-60)^2+(55-60)^2+(61-60)^2+(63-60)^2+(54-60)^2$ $=7^2+(-5)^2+1^2+3^2+(-6)^2$ $=120$
(总体)方差	$\frac{1834}{5}=366.8$	$\frac{120}{5}=24$
(总体)标准差	$\sqrt{366.8}=19.2$	$\sqrt{24}=4.9$

注：在方差中，也有不采用“数据个数”而采用“数据个数 -1”作为分母的情况，我们将其称为样本方差。由于篇幅所限，这两种方差的区别在本书中就不做讨论了。

✿ 9. 概率密度函数 ✿

■9.1 χ^2 分布

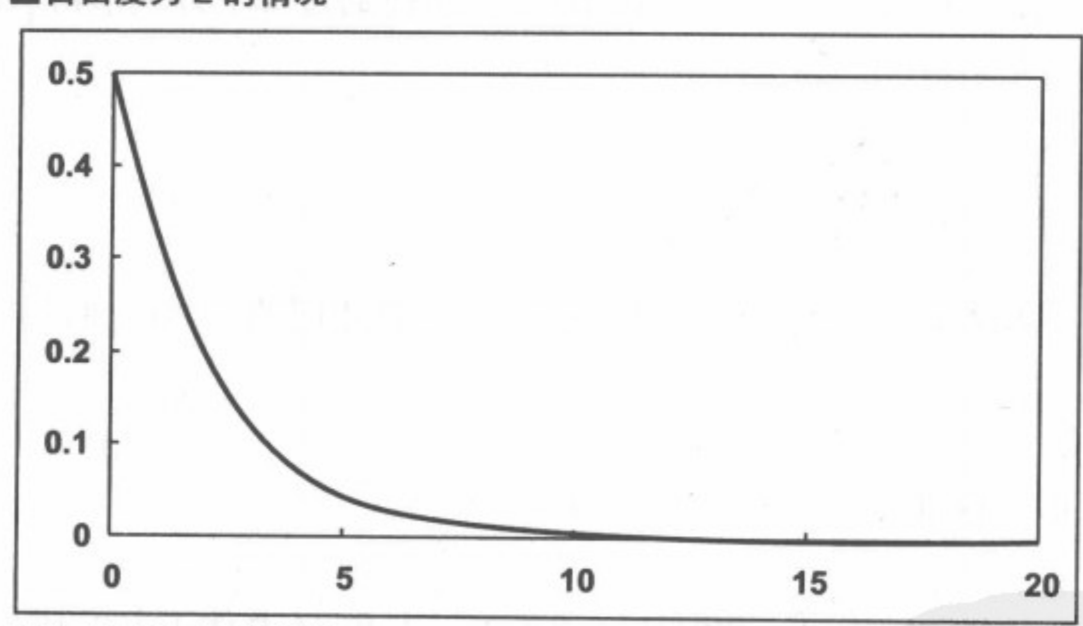
在统计学中，常常会出现下面介绍的概率密度函数。

$$f(x)=\begin{cases}\dfrac{1}{2^{\frac{\text{自由度}}{2}}\times\int_0^{\infty}x^{\frac{\text{自由度}}{2}-1}e^{-x}\,dx}\times x^{\frac{\text{自由度}}{2}-1}\times e^{-\frac{x}{2}} & x>0\text{时,}\\ 0 & x\leqslant 0\text{时}\end{cases}$$

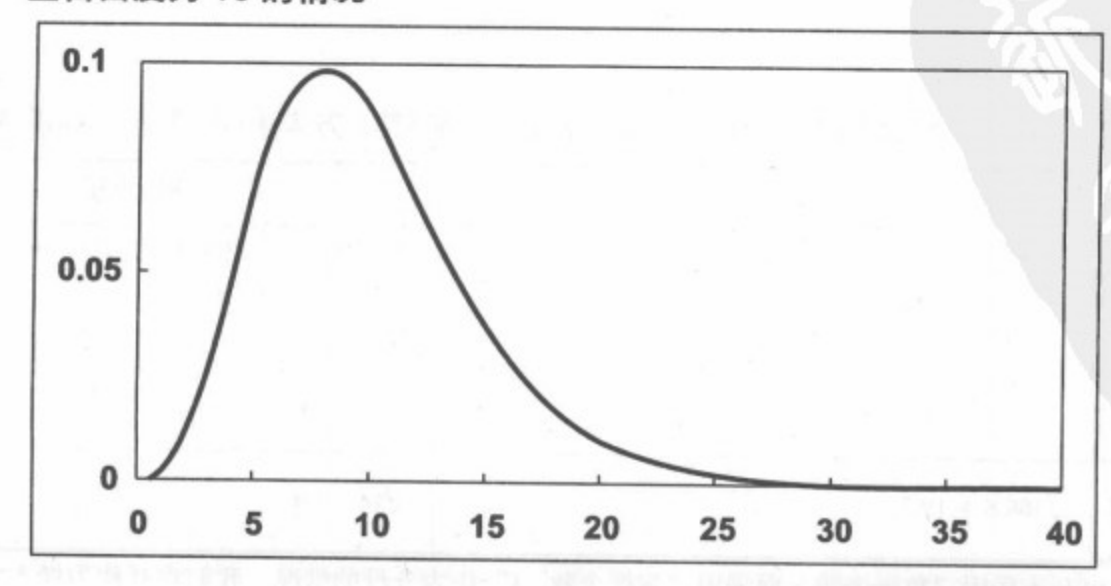
如果 x 的概率密度函数满足上述条件的话，在统计学中我们就将其表述为“x 服从自由度为○○的 χ^2分布”。

“自由度是什么？”，这或许会让读者摸不着头脑。其实“自由度是什么”这个问题就相当于在问“一次函数 $f(x)=ax+b$ 中的 a 是什么”。所谓自由度,不过就是“影响函数图像形状的一个值”而已，除此之外便没有其他意义了。

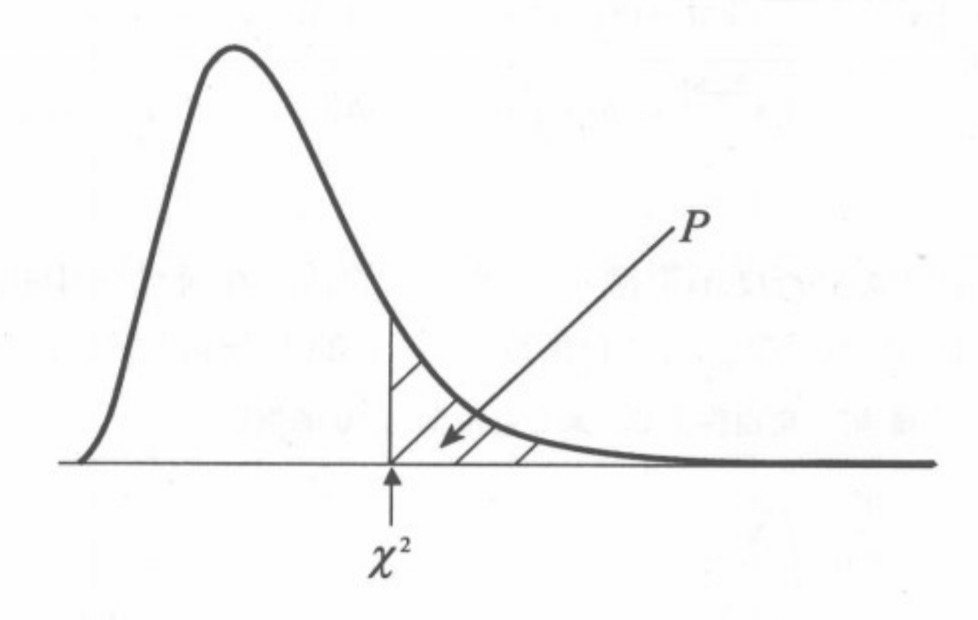
■自由度为 2 的情况



■自由度为 10 的情况



在实际应用中，存在一种叫做"χ^2分布表"的表格。使用这个表格，我们就可将与下图中斜线部分的概率（= 面积）P 相对应的横轴坐标查找出来。图中的 χ^2 读做"卡方"。



下面是我们节选的一部分 χ^2 分布表。

◆表 1.7　χ^2 分布表

自由度 \ P	0.995	0.99	0.975	0.95	0.05	0.025	0.01	0.005
1	0.000039	0.0002	0.0010	0.0039	3.8415	5.0239	6.6349	7.8794
2	0.0100	0.0201	0.0506	0.1026	5.9915	7.3778	9.2104	10.5965
3	0.0717	0.1148	0.2158	0.3518	7.8147	9.3484	11.3449	12.8381
4	0.2070	0.2971	0.4844	0.7107	9.4877	11.1433	13.2767	14.8602
5	0.4118	0.5543	0.8312	1.1455	11.0705	12.8325	15.0863	16.7496
6	0.6757	0.8721	1.2373	1.6354	12.5916	14.4494	16.8119	18.5475
7	0.9893	1.2390	1.6899	2.1673	14.0671	16.0128	18.4753	20.2777
8	1.3444	1.6465	2.1797	2.7326	15.5073	17.5345	20.0902	21.9549
9	1.7349	2.0879	2.7004	3.3251	16.9190	19.0228	21.6660	23.5893
10	2.1558	2.5582	3.2470	3.9403	18.3070	20.4832	23.2093	25.1881
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

例

当 P 为 0.05、自由度为 2 的时候，χ^2 的值为 5.9915。

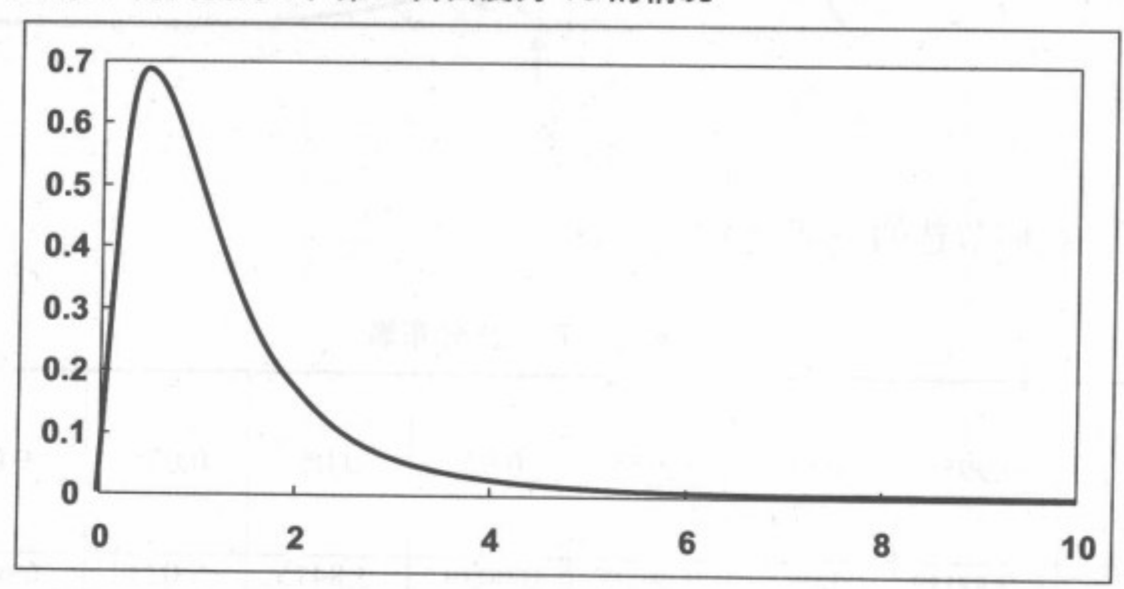
■9.2 *F*分布

在统计学中，常常会出现下面介绍的这种概率密度函数。

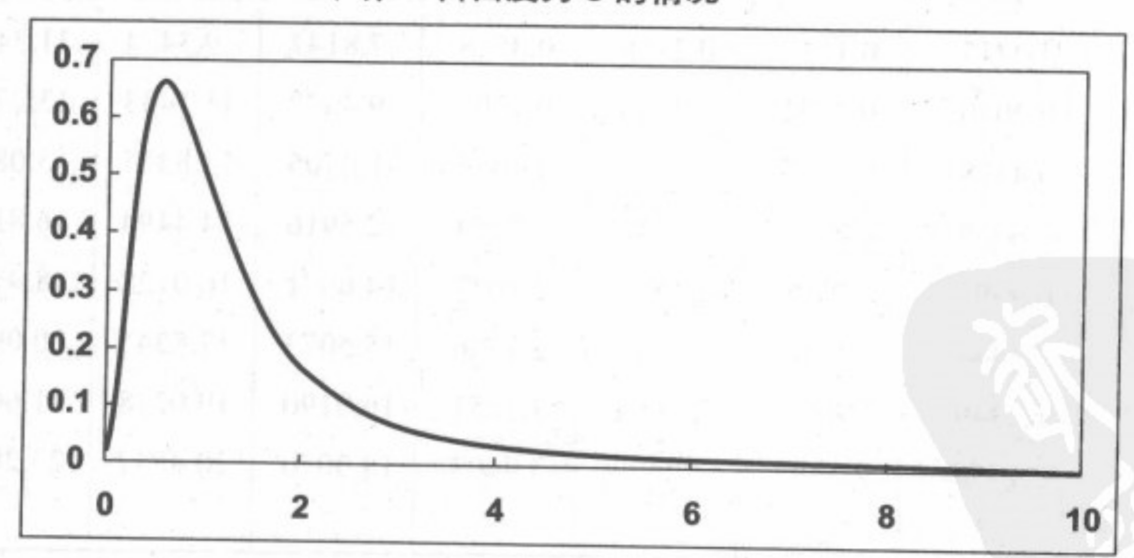
$$f(x)=\begin{cases}\dfrac{\left(\int_0^{\infty}x^{\frac{\text{第1自由度}+\text{第2自由度}}{2}-1}e^{-x}dx\right)\times(\text{第1自由度})^{\frac{\text{第1自由度}}{2}}\times(\text{第2自由度})^{\frac{\text{第2自由度}}{2}}}{\left(\int_0^{\infty}x^{\frac{\text{第1自由度}}{2}-1}e^{-x}dx\right)\times\left(\int_0^{\infty}x^{\frac{\text{第2自由度}}{2}-1}e^{-x}dx\right)}\times\dfrac{x^{\frac{\text{第1自由度}}{2}-1}}{(\text{第1自由度})\times x+(\text{第2自由度})^{\frac{\text{第1自由度}+\text{第2自由度}}{2}}} & x>0\text{时，}\\ 0 & x\leqslant 0\text{时}\end{cases}$$

如果 x 的概率密度函数满足上述条件的话，在统计学中我们就将其表述为“x 服从第一自由度为○○、第 2 自由度为 ×× 的 F 分布”。

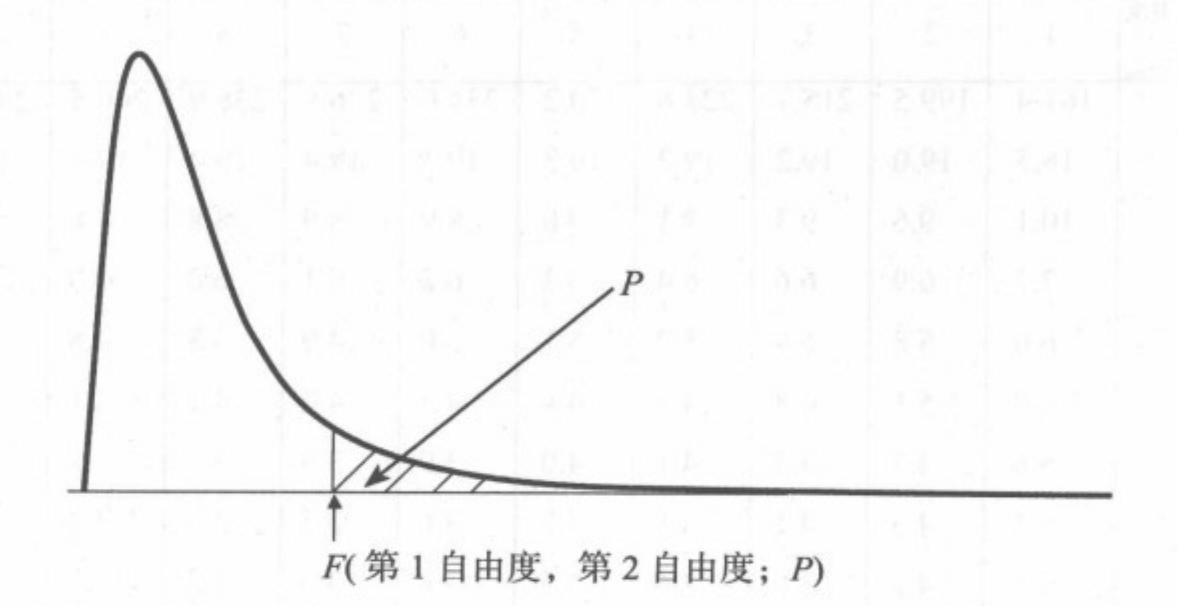
■ 第 1 自由度为 5、第 2 自由度为 10 的情况



■ 第 1 自由度为 10、第 2 自由度为 5 的情况



在实际应用中，存在一种叫做“F 分布表”的表格。使用这个表格，我们就可将与下图中斜线部分的概率（=面积）P 相对应的横轴坐标查找出来。



下面是我们节选的一部分 F 分布表：

◆表 1.8　P 为 0.05 时的 F 分布表

第2自由度＼第1自由度	1	2	3	4	5	6	7	8	9	10	…
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	…
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	…
3	10.1	9.6	9.3	9.1	9.0	8.9	8.9	8.8	8.8	8.8	…
4	7.7	6.9	6.6	6.4	6.3	6.2	6.1	6.0	6.0	6.0	…
5	6.6	5.8	5.4	5.2	5.1	5.0	4.9	4.8	4.8	4.7	…
6	6.0	5.1	4.8	4.5	4.4	4.3	4.2	4.1	4.1	4.1	…
7	5.6	4.7	4.3	4.1	4.0	3.9	3.8	3.7	3.7	3.6	…
8	5.3	4.5	4.1	3.8	3.7	3.6	3.5	3.4	3.4	3.3	…
9	5.1	4.3	3.9	3.6	3.5	3.4	3.3	3.2	3.2	3.1	…
10	5.0	4.1	3.7	3.5	3.3	3.2	3.1	3.1	3.0	3.0	…
11	4.8	4.0	3.6	3.4	3.2	3.1	3.0	2.9	2.9	2.9	…
12	4.7	3.9	3.5	3.3	3.1	3.0	2.9	2.8	2.8	2.8	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

◆表 1.9　P 为 0.01 时的 F 分布表

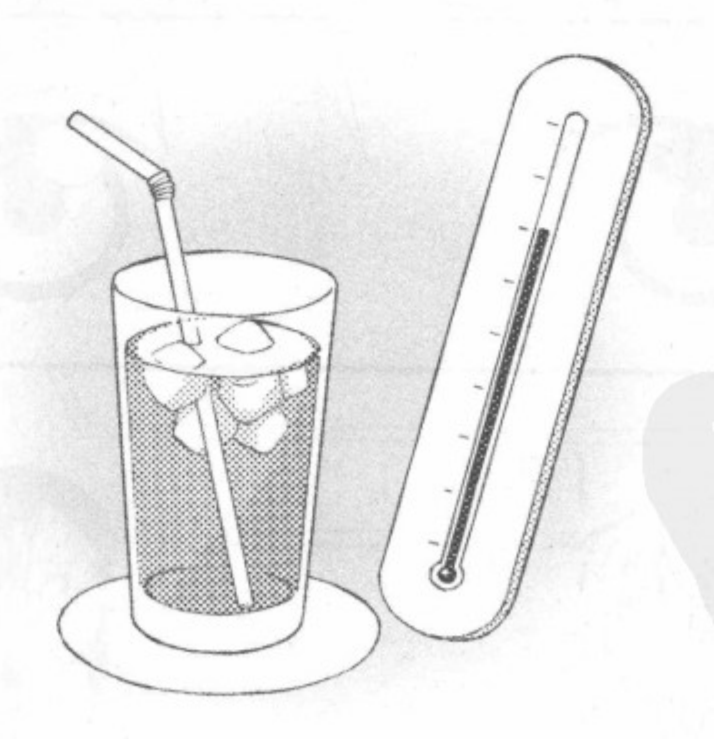
第2自由度＼第1自由度	1	2	3	4	5	6	7	8	9	10	…
1	4052.2	4999.3	5403.5	5624.3	5764.0	5859.0	5928.3	5981.0	6022.4	6055.9	…
2	98.5	99.0	99.2	99.3	99.3	99.3	99.4	99.4	99.4	99.4	…
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	…
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	…
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	…
6	13.7	10.9	9.8	9.1	8.7	8.5	8.3	8.1	8.0	7.9	…
7	12.2	9.5	8.5	7.8	7.5	7.2	7.0	6.8	6.7	6.6	…
8	11.3	8.6	7.6	7.0	6.6	6.4	6.2	6.0	5.9	5.8	…
9	10.6	8.0	7.0	6.4	6.1	5.8	5.6	5.5	5.4	5.3	…
10	10.0	7.6	6.6	6.0	5.6	5.4	5.2	5.1	4.9	4.8	…
11	9.6	7.2	6.2	5.7	5.3	5.1	4.9	4.7	4.6	4.5	…
12	9.3	6.9	6.0	5.4	5.1	4.8	4.6	4.5	4.4	4.3	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

例

当 P 为 0.05、第 1 自由度为 1、第 2 自由度为 12 的时候，也可以记作 F（第 1 自由度，第 2 自由度；P），即 $F(1，12；0.05)$ 的值为 4.7。

◆第2章◆

回归分析





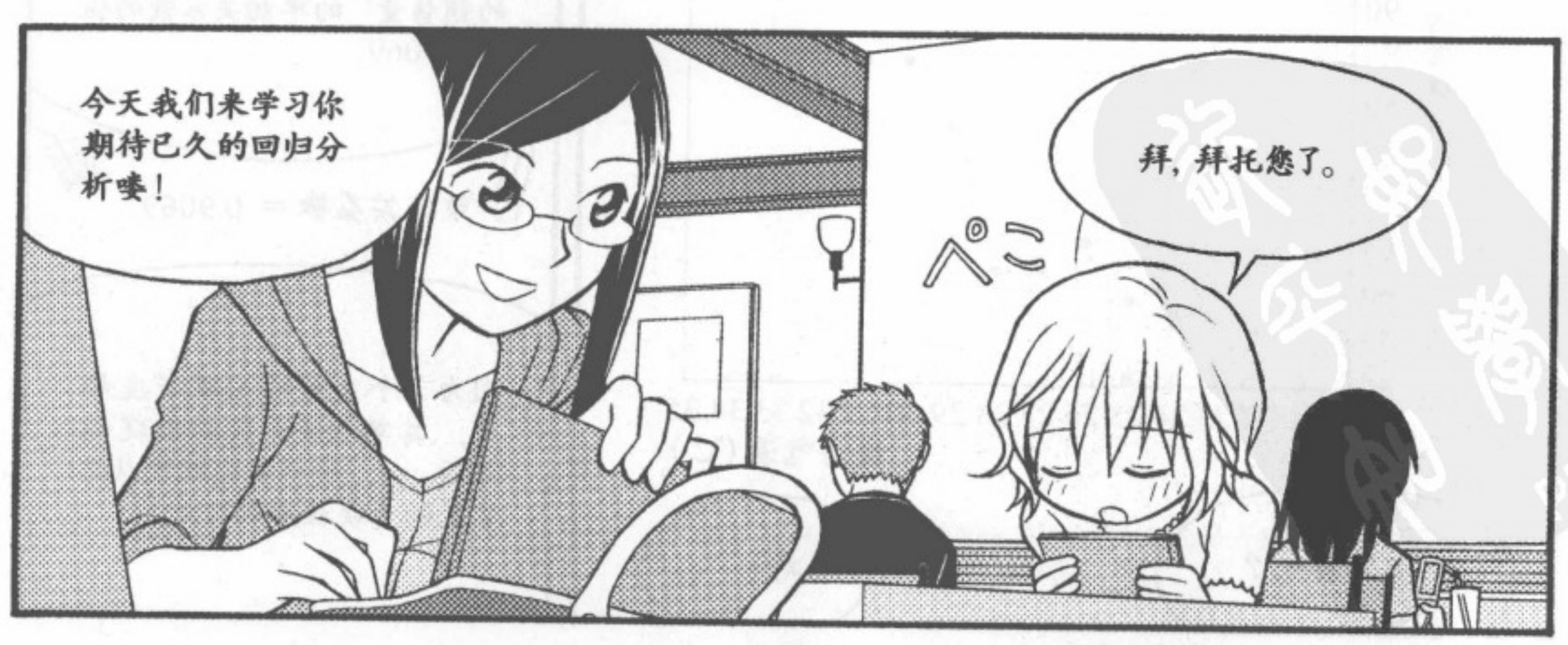
1. 回归分析
哈！回归分析……
就是这样做对吧？
是啊！
小美羽也了解啊！
ぼー…
美羽！
はっ
你没事吧？

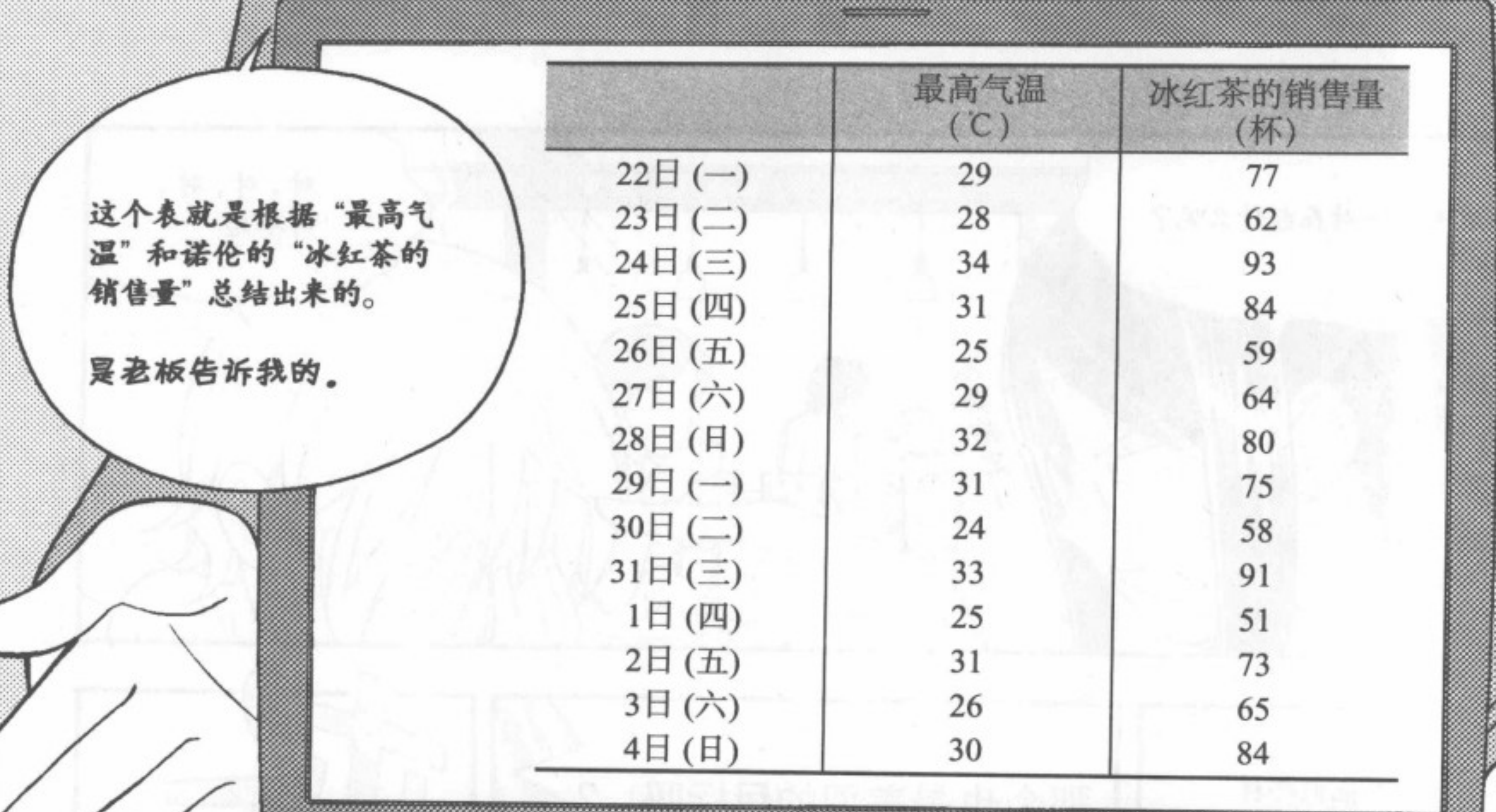

看着那一对在想什么呢？
从实招来！
对，对，对，
对不起……


可以像他们那样
学习……
该多好呀！


那个也是美羽的目标吗！？
对不起！

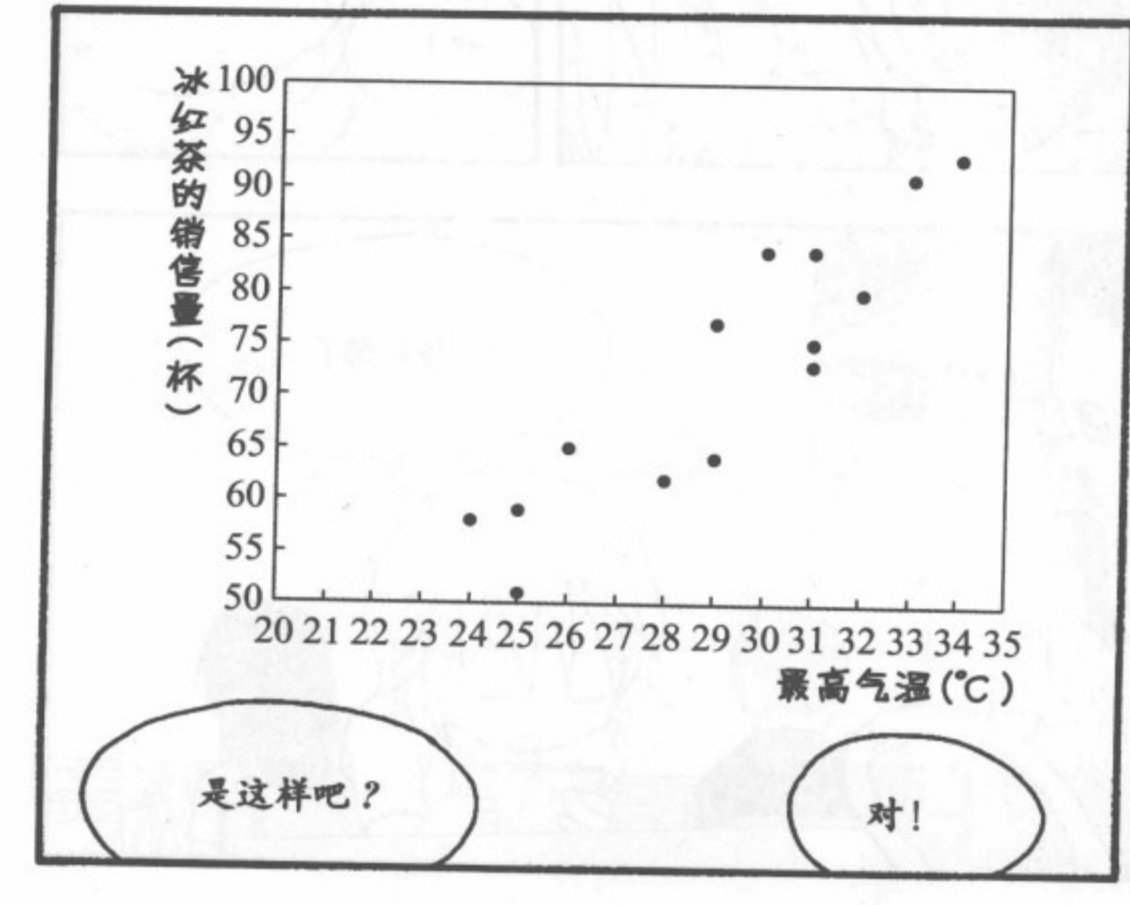

ポー
喂！


今天我们来学习你
期待已久的回归分
析喽！
拜，拜托您了。
ぺこ



	最高气温（℃）	冰红茶的销售量（杯）
22日 (一)	29	77
23日 (二)	28	62
24日 (三)	34	93
25日 (四)	31	84
26日 (五)	25	59
27日 (六)	29	64
28日 (日)	32	80
29日 (一)	31	75
30日 (二)	24	58
31日 (三)	33	91
1日 (四)	25	51
2日 (五)	31	73
3日 (六)	26	65
4日 (日)	30	84





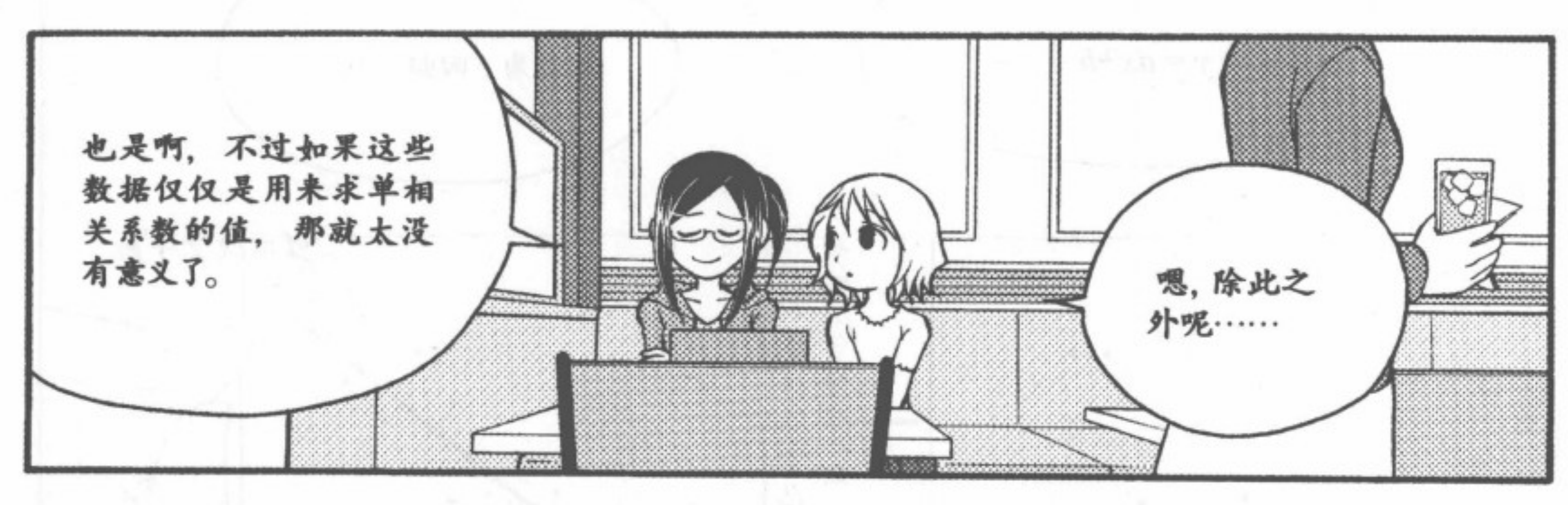
我们先不讲如何计算，只须知道“最高气温”和“冰红茶的销售量”的单相关系数的值是0.9069。

单相关系数 = 0.9069

因为2个变量的关联程度越大，其单相关系数的值就越接近 ±1，所以我们就可以说它们的关联程度相当大。

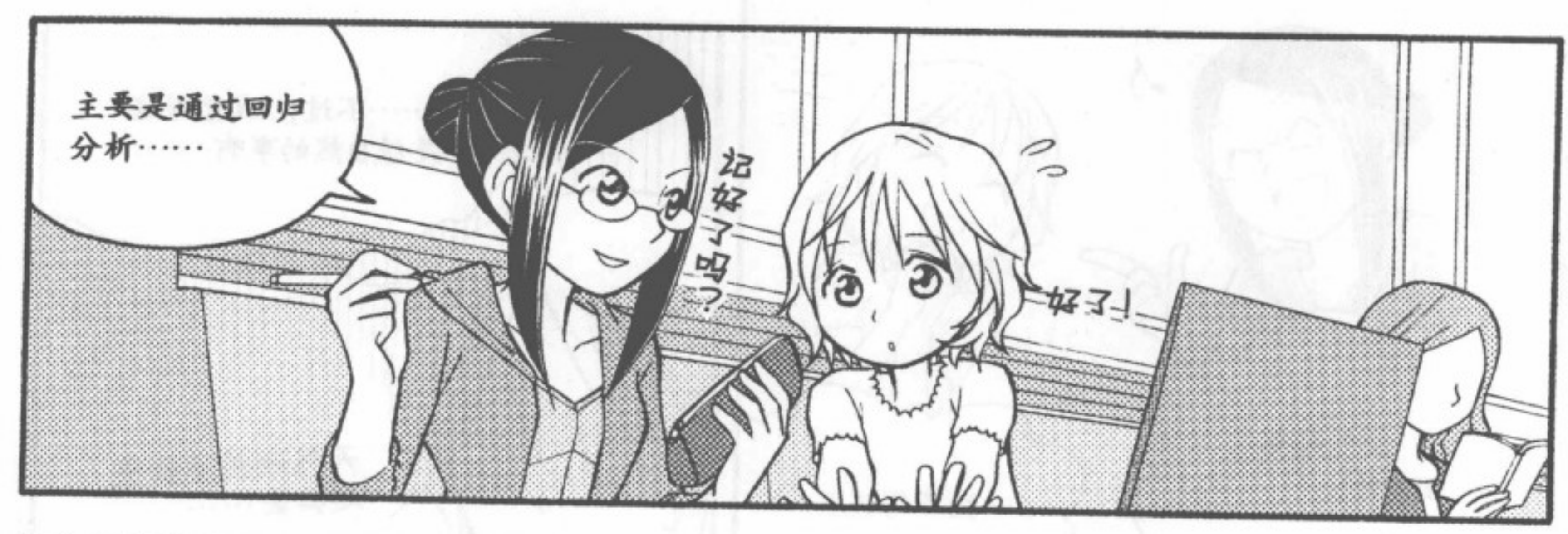

所以，天气越热
越好卖哟！
怪不得呢！

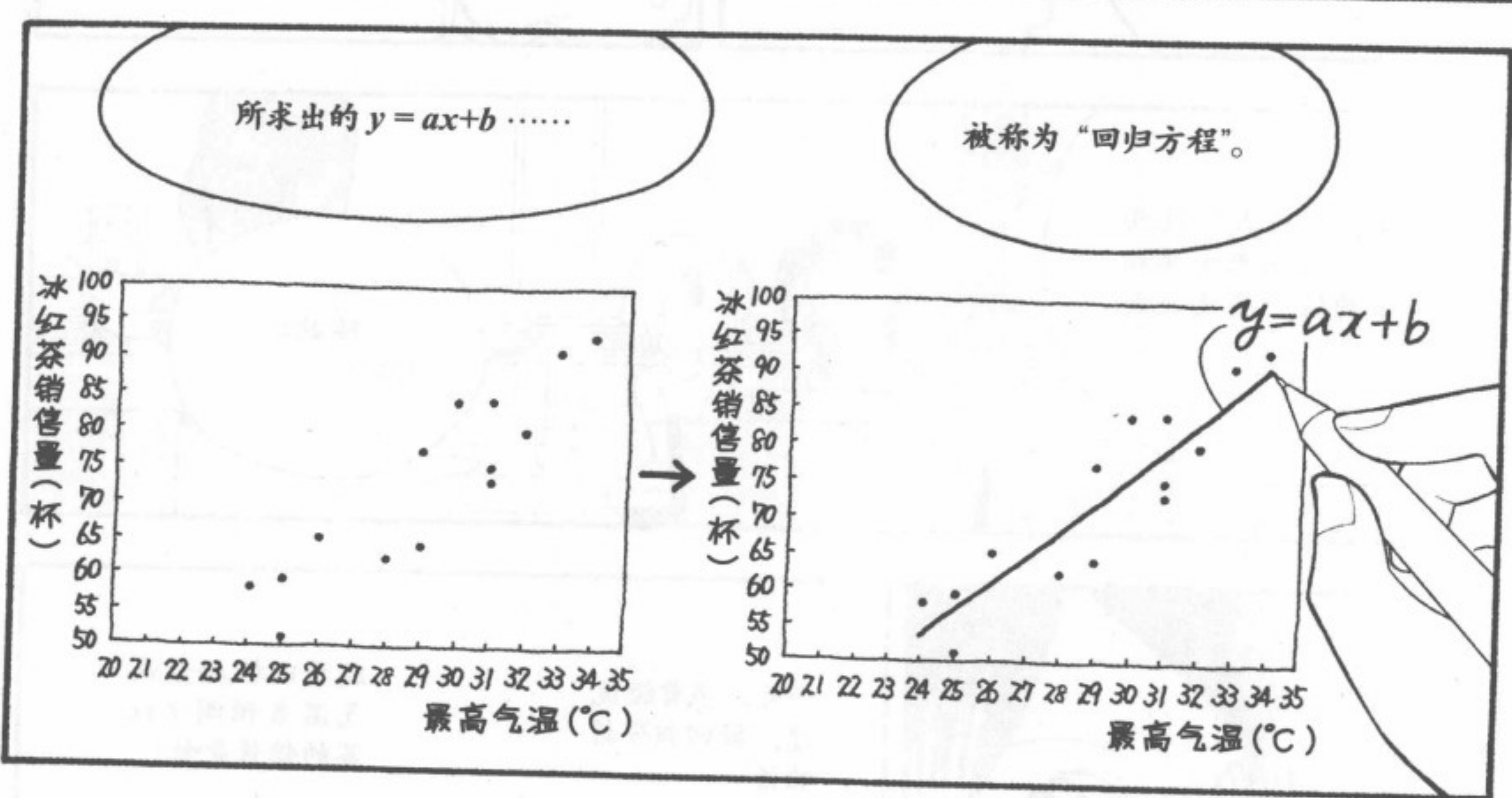

……不过，我觉得这
是很自然的事啊……
天气越热冰红茶
越好卖……

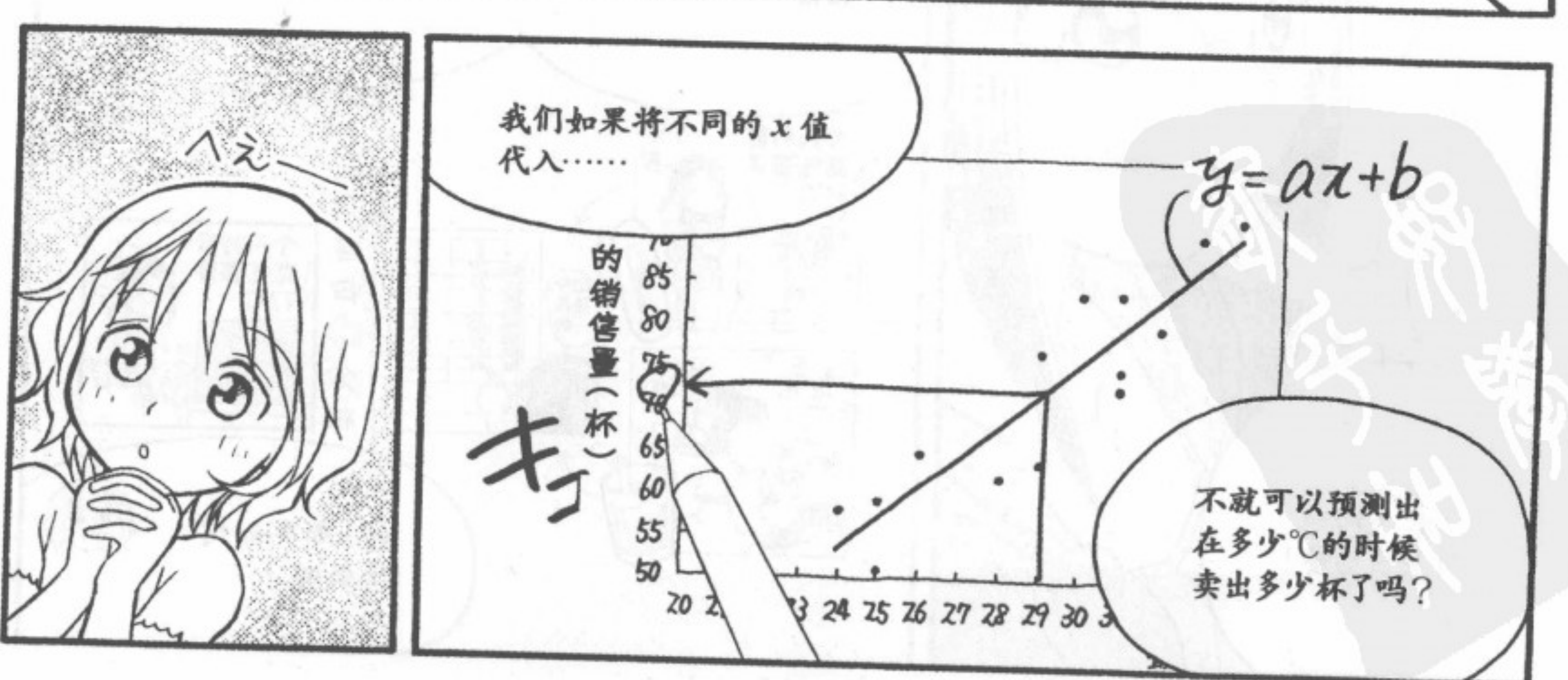

也是啊，不过如果这些
数据仅仅是用来求单相
关系数的值，那就太没
有意义了。
嗯，除此之
外呢……

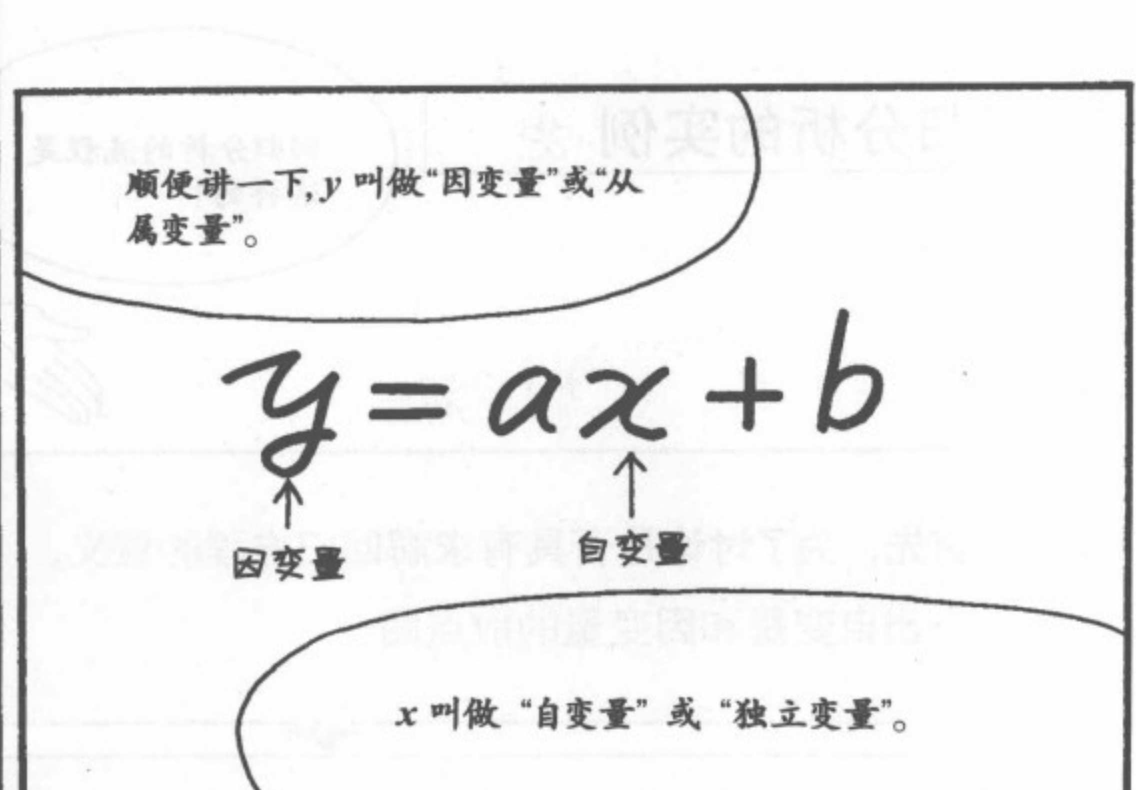
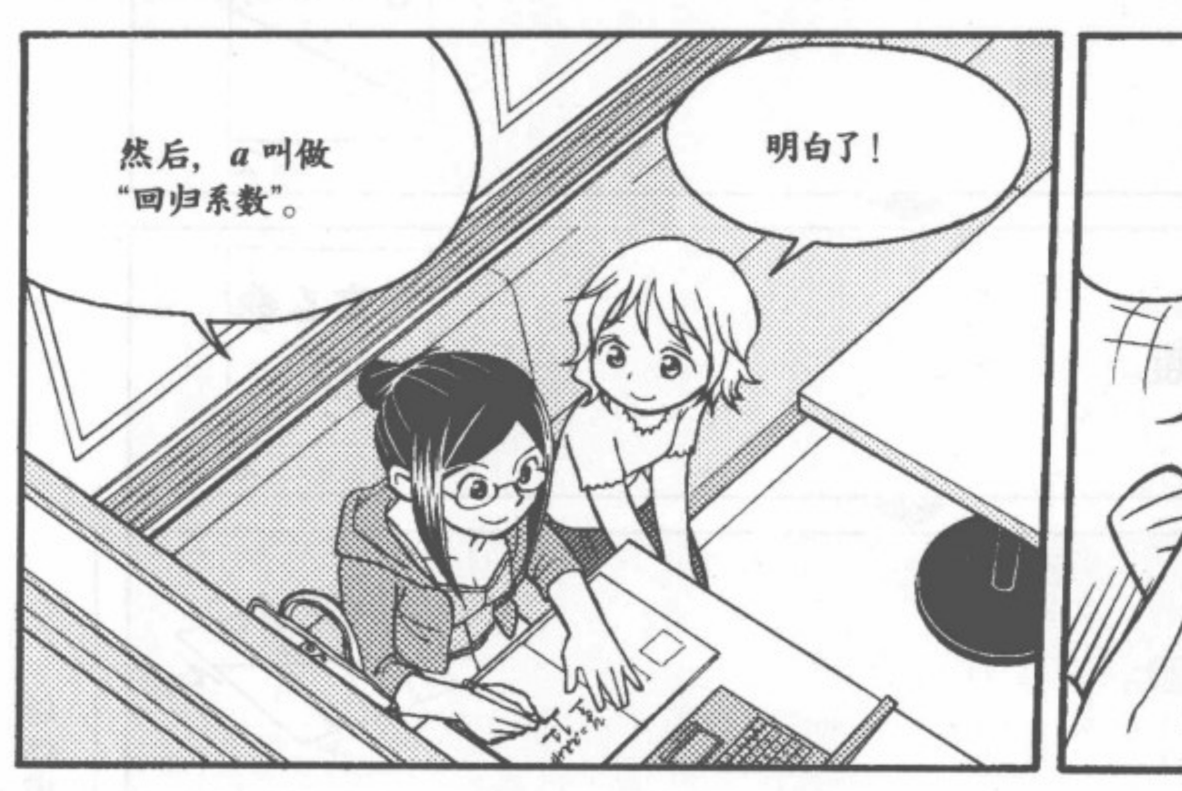

接下来，回归
分析要出场喽！

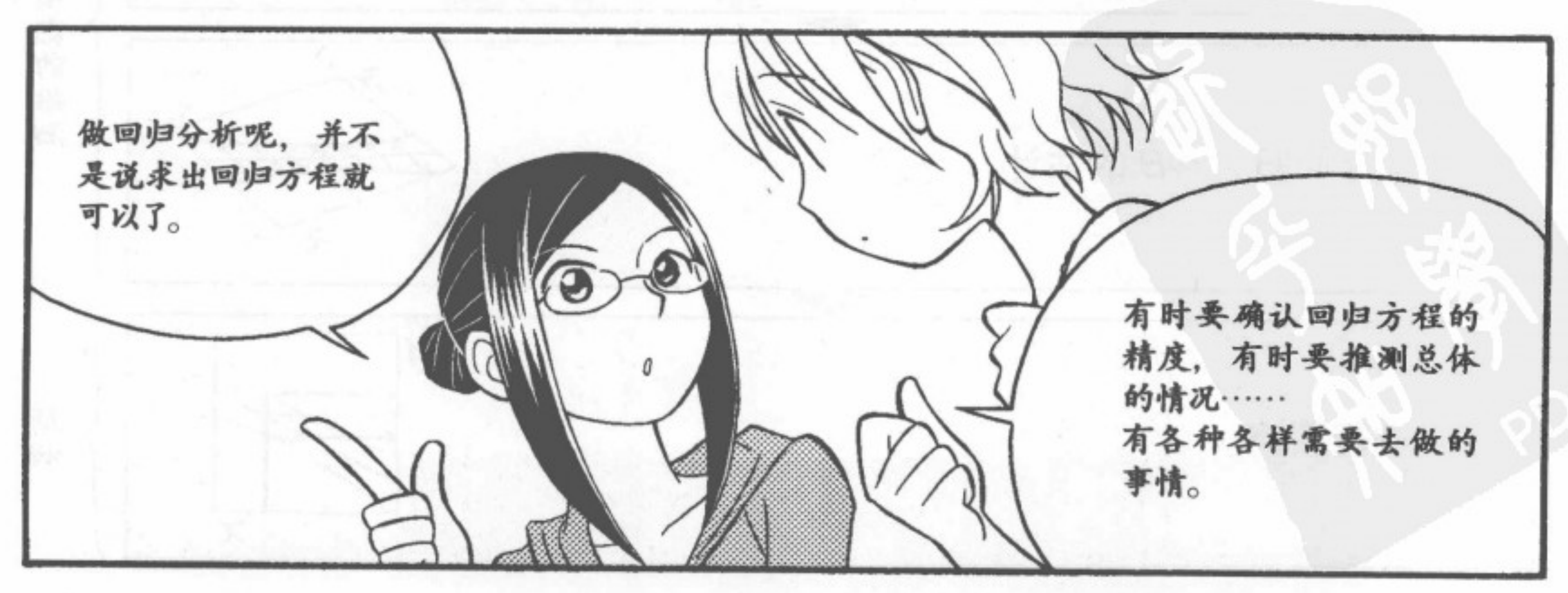

那天，我曾经说
过，做回归分析
的话……
就可以根据最高
气温来预测冰红
茶的销售量啦！
今天的最高气温是31°C
冰红茶
冰红茶
记录今天的
回归分析
今天的最高气温是27°C
シャキーン
哇！那要如何
去做呢……


主要是通过回归分析……
记好了吗？
好了！

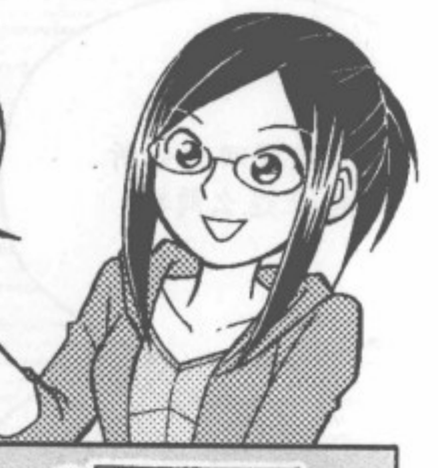

所求出的 y = ax+b ……
被称为“回归方程”。
冰红茶销售量（杯）
最高气温(℃)
y=ax+b


へえ——
我们如果将不同的 x 值代入……
y=ax+b
的销售量（杯）
キュ
不就可以预测出在多少℃的时候卖出多少杯了吗？

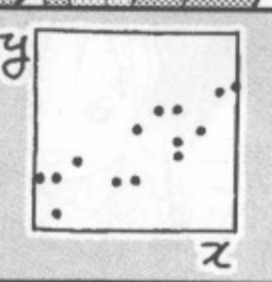

怎么觉得回归分析看起来这么简单。
フフフ……

顺便讲一下，y 叫做“因变量”或“从属变量”。
y = ax + b
因变量
自变量
x 叫做“自变量”或“独立变量”。

然后，a 叫做“回归系数”。
明白了！

那么，请尽快教我回归方程的求解方法吧！
等一下！

做回归分析呢，并不是说求出回归方程就可以了。
有时要确认回归方程的精度，有时要推测总体的情况……
有各种各样需要去做的事情。

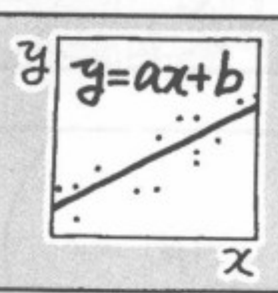
✿ 2. 回归分析的实例 ✿



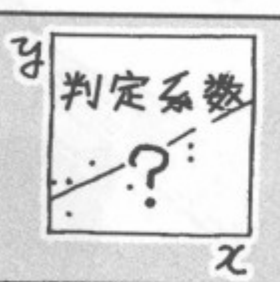
① 首先，为了讨论是否具有求解回归方程的意义，画出自变量和因变量的散点图。



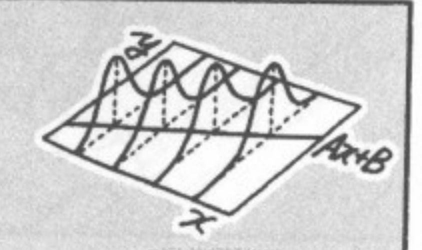
② 求解回归方程。



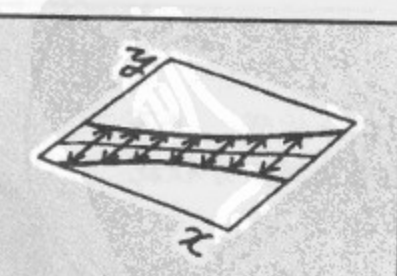
③ 确认回归方程的精度。



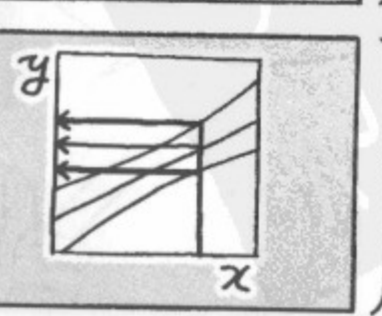
④ 进行回归系数的检验。



⑤ 总体回归 *Ax+B* 的估计。

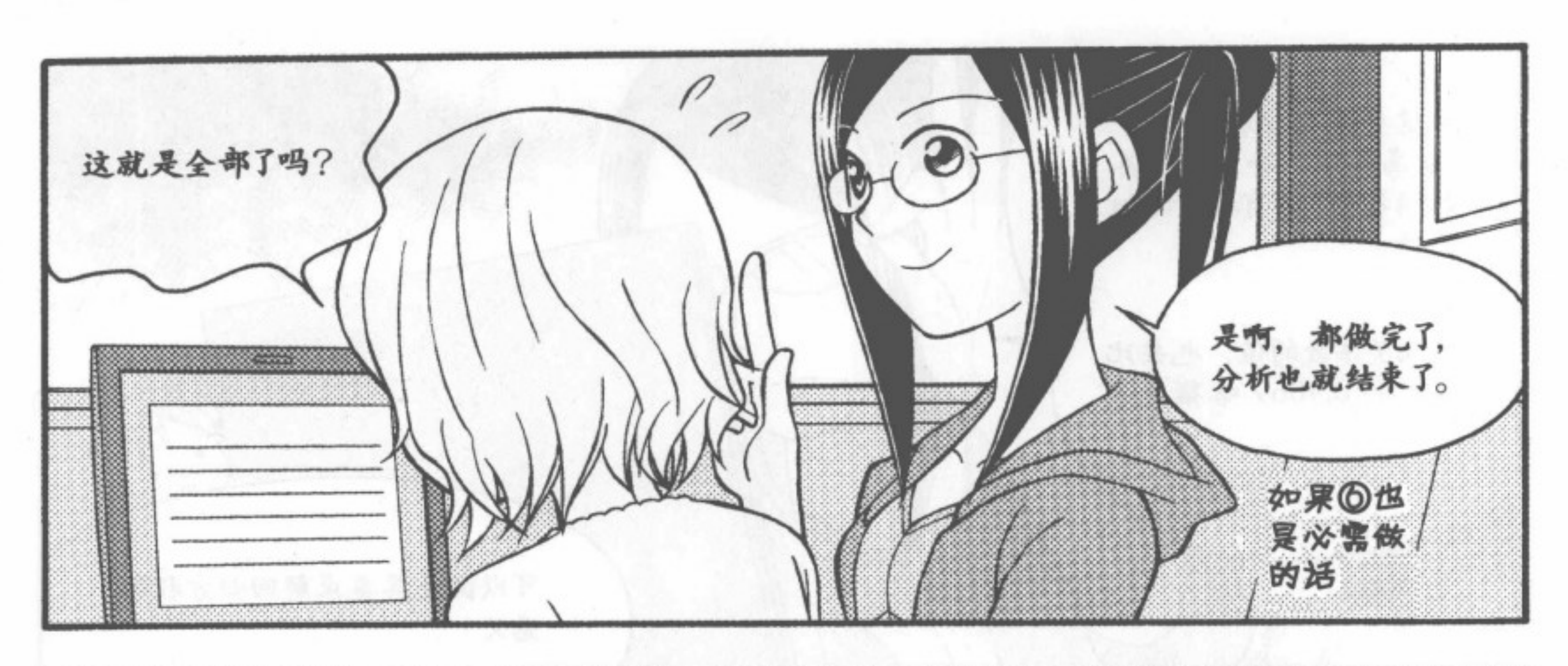


⑥ 进行预测。



总体情况的推测（④⑤）

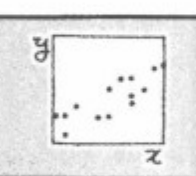
预测（⑥）

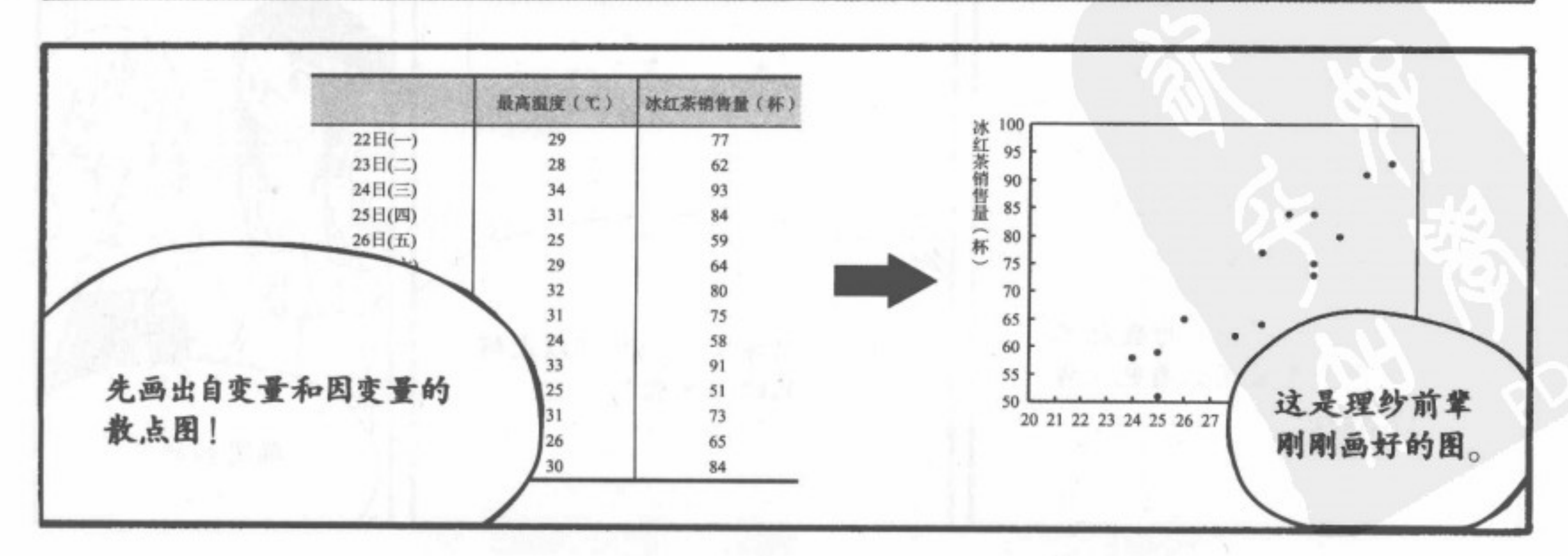






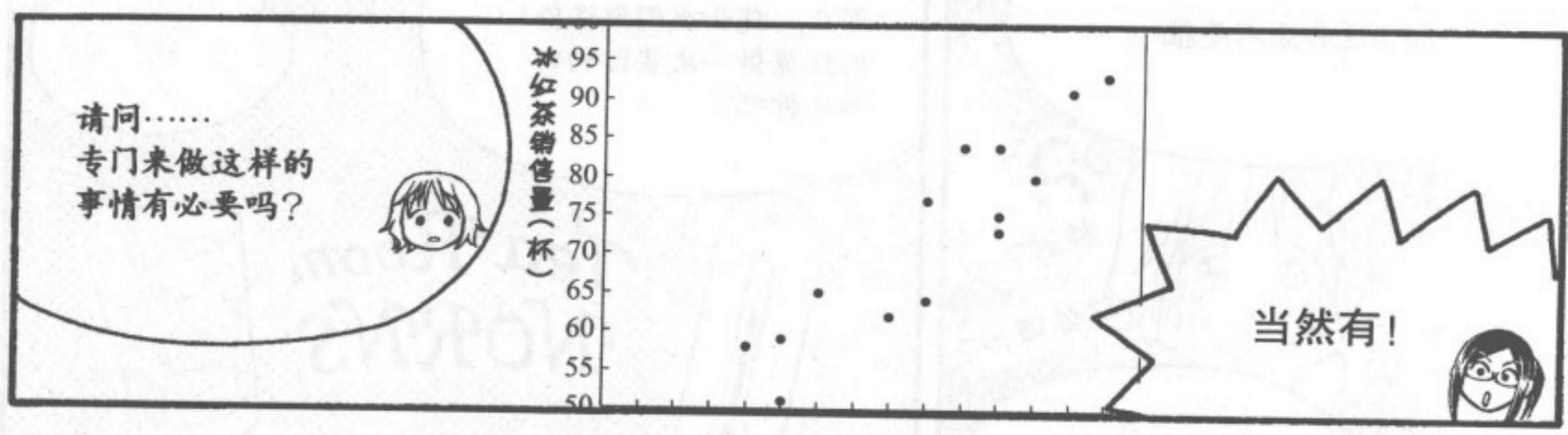
① 首先，为了讨论是否具有求解回归方程的意义，画出自变量和因变量的散点图

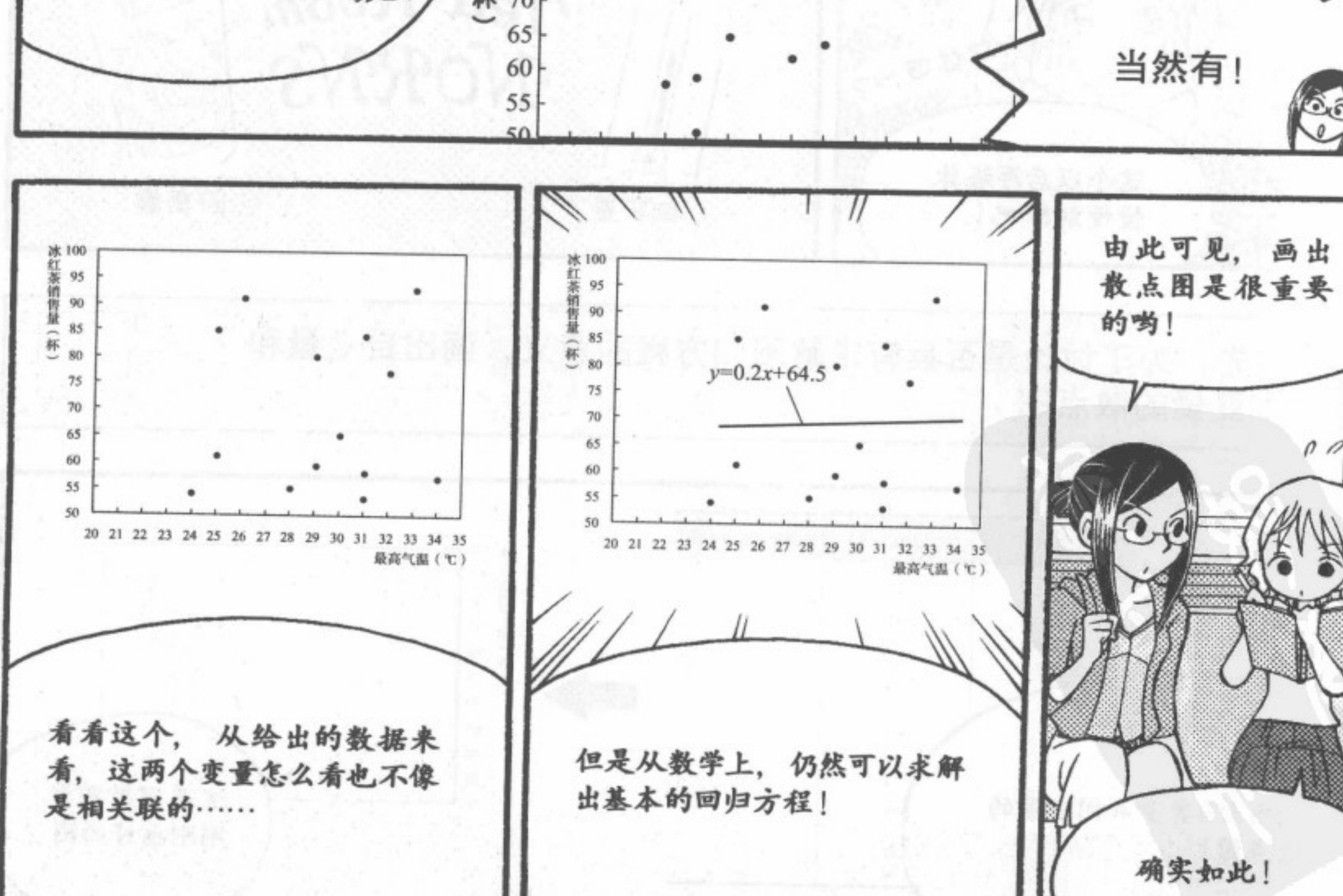




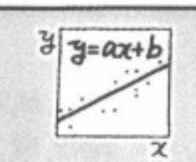
	最高温度（℃）	冰红茶销售量（杯）
22日(一)	29	77
23日(二)	28	62
24日(三)	34	93
25日(四)	31	84
26日(五)	25	59
	29	64
	32	80
	31	75
	24	58
	33	91
	25	51
	31	73
	26	65
	30	84

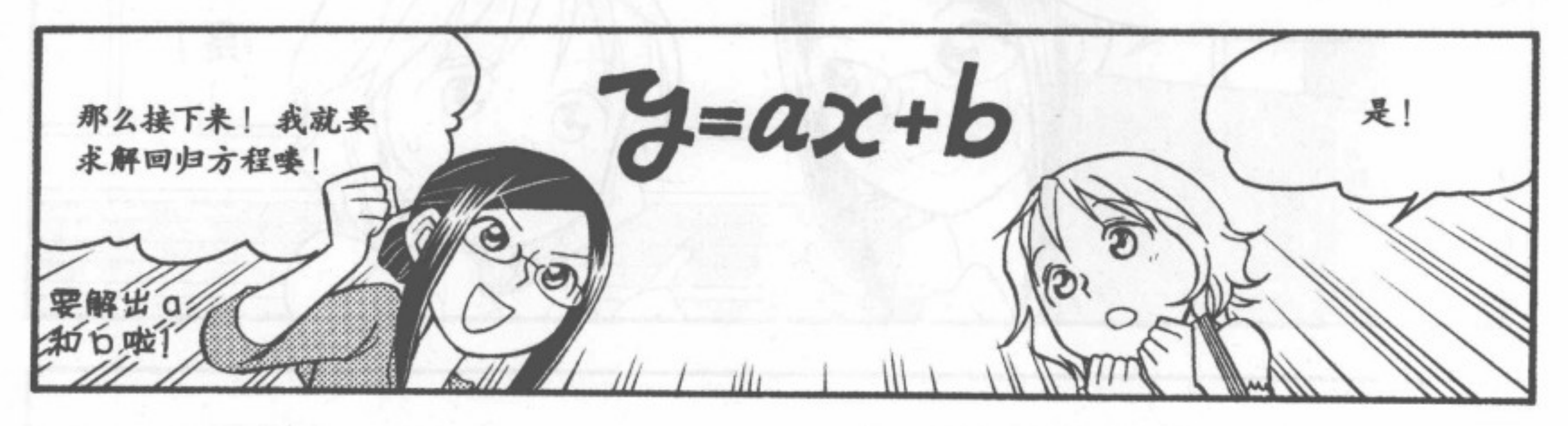

从这些数据看来……
“最高气温”和“冰红茶的销售量”具有很强的相关性。
单相关系数的值，也要比实际的0.9069略微大一些……
可以说是具备求解回归方程的意义！


请问……
专门来做这样的事情有必要吗？
冰红茶销售量（杯）
当然有！

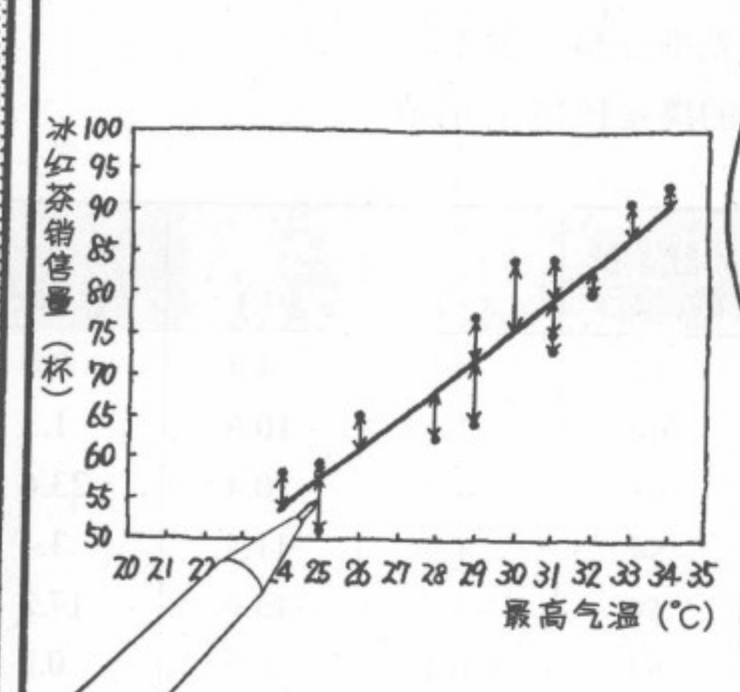

冰红茶销售量（杯）
最高气温（℃）
看看这个，从给出的数据来看，这两个变量怎么看也不像是相关联的……
$y=0.2x+64.5$
但是从数学上，仍然可以求解出基本的回归方程！
由此可见，画出散点图是很重要的哟！
确实如此！

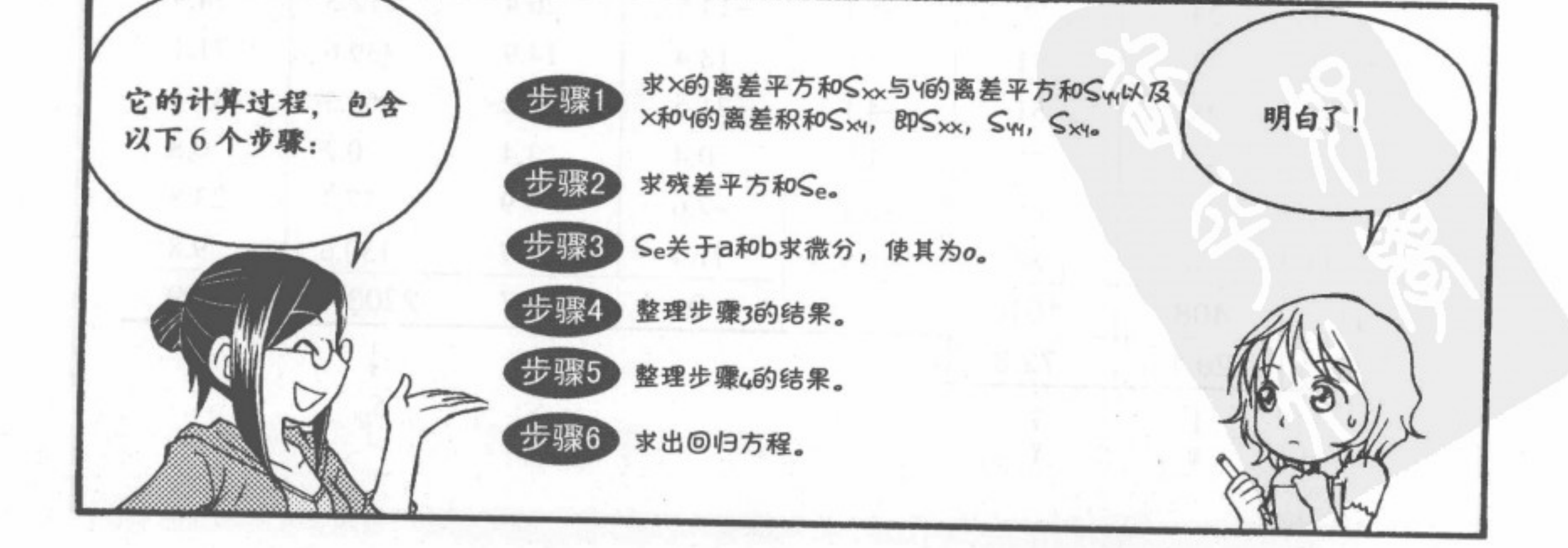
② 求解回归方程

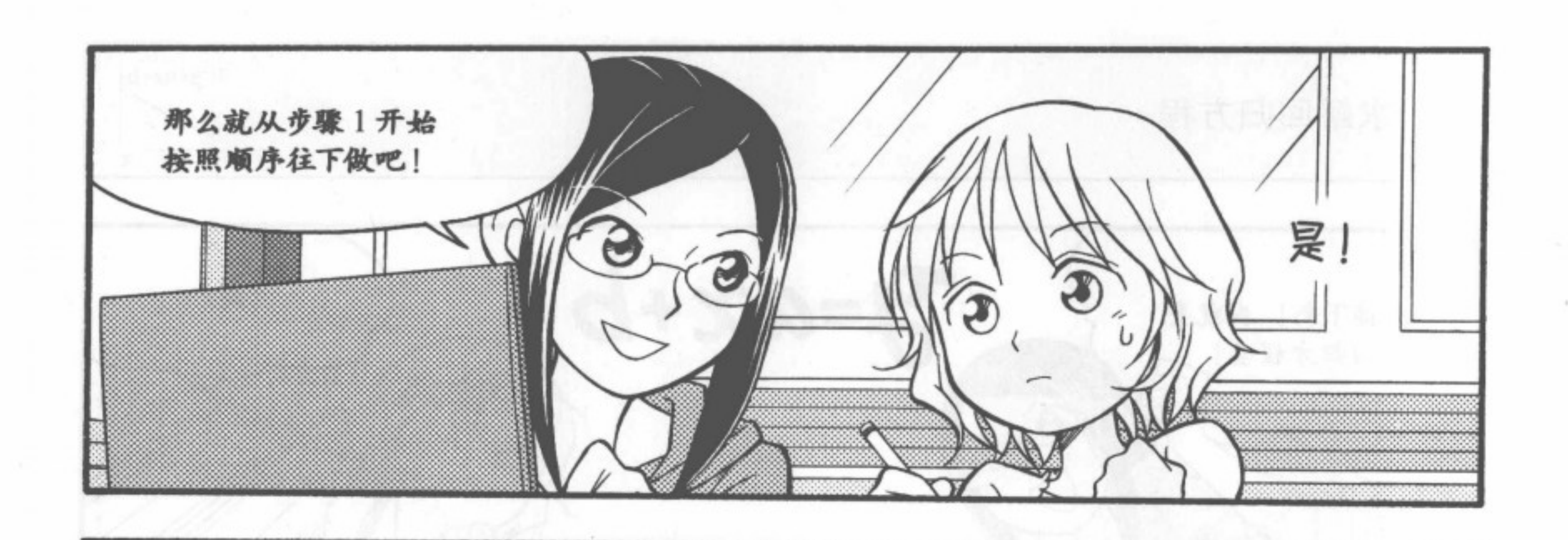

y
y=ax+b
x


那么接下来！我就要求解回归方程喽！
y=ax+b
要解出a和b啦！
是！


具体思路是这样的……


冰红茶销售量（杯）
100
95
90
85
80
75
70
65
60
55
50
20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35
最高气温（℃）
在图中这些竖线长度的平方之和达到最小时，求出 a 和 b。
这种思考方法就是所谓的"最小二乘法"。


它的计算过程，包含以下6个步骤：
步骤1 求x的离差平方和Sxx与y的离差平方和Syy以及x和y的离差积和Sxy，即Sxx，Syy，Sxy。
步骤2 求残差平方和Se。
步骤3 Se关于a和b求微分，使其为0。
步骤4 整理步骤3的结果。
步骤5 整理步骤4的结果。
步骤6 求出回归方程。
明白了！



步骤 1
- 求出 x 的离差平方和 S_{xx} 的值
- 求出 y 的离差平方和 S_{yy} 的值
- 求出 x 和 y 的离差积和 S_{xy} 的值

	最高气温 x	冰红茶的销售量 y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$
22日(一)	29	77	−0.1	4.4	0.0	19.6	−0.6
23日(二)	28	62	−1.1	−10.6	1.3	111.8	12.1
24日(三)	34	93	4.9	20.4	23.6	417.3	99.2
25日(四)	31	84	1.9	11.4	3.4	130.6	21.2
26日(五)	25	59	−4.1	−13.6	17.2	184.2	56.2
27日(六)	29	64	−0.1	−8.6	0.0	73.5	1.2
28日(日)	32	80	2.9	7.4	8.2	55.2	21.2
29日(一)	31	75	1.9	2.4	3.4	5.9	4.5
30日(二)	24	58	−5.1	−14.6	26.4	212.3	74.9
31日(三)	33	91	3.9	18.4	14.9	339.6	71.1
1日(四)	25	51	−4.1	−21.6	17.2	465.3	89.4
2日(五)	31	73	1.9	0.4	3.4	0.2	0.8
3日(六)	26	65	−3.1	−7.6	9.9	57.3	23.8
4日(日)	30	84	0.9	11.4	0.7	130.6	9.8
总计	408	1016	0	0	129.7	2203.4	484.9
平均	29.1	72.6					
	↓ $\bar{x}$	↓ $\bar{y}$			↓ S_{xx}	↓ S_{yy}	↓ S_{xy}

步骤 2 下面给出的是具体的运算过程

表中“y”叫做实测值。

“$\hat{y}=ax+b$”叫做预测值。

“$y-\hat{y}$”叫做残差，一般用“e”来表示。

	最高气温 x	冰红茶的销售量y	冰红茶的销售量 $\hat{y}=ax+b$	$y-\hat{y}$	$(y-\hat{y})^2$
22日 (一)	29	77	$a\times 29+b$	$77-(a\times 29+b)$	$[77-(a\times 29+b)]^2$
23日 (二)	28	62	$a\times 28+b$	$62-(a\times 28+b)$	$[62-(a\times 28+b)]^2$
24日 (三)	34	93	$a\times 34+b$	$93-(a\times 34+b)$	$[93-(a\times 34+b)]^2$
25日 (四)	31	84	$a\times 31+b$	$84-(a\times 31+b)$	$[84-(a\times 31+b)]^2$
26日 (五)	25	59	$a\times 25+b$	$59-(a\times 25+b)$	$[59-(a\times 25+b)]^2$
27日 (六)	29	64	$a\times 29+b$	$64-(a\times 29+b)$	$[64-(a\times 29+b)]^2$
28日 (日)	32	80	$a\times 32+b$	$80-(a\times 32+b)$	$[80-(a\times 32+b)]^2$
29日 (一)	31	75	$a\times 31+b$	$75-(a\times 31+b)$	$[75-(a\times 31+b)]^2$
30日 (二)	24	58	$a\times 24+b$	$58-(a\times 24+b)$	$[58-(a\times 24+b)]^2$
31日 (三)	33	91	$a\times 33+b$	$91-(a\times 33+b)$	$[91-(a\times 33+b)]^2$
1日 (四)	25	51	$a\times 25+b$	$51-(a\times 25+b)$	$[51-(a\times 25+b)]^2$
2日 (五)	31	73	$a\times 31+b$	$73-(a\times 31+b)$	$[73-(a\times 31+b)]^2$
3日 (六)	26	65	$a\times 26+b$	$65-(a\times 26+b)$	$[65-(a\times 26+b)]^2$
4日 (日)	30	84	$a\times 30+b$	$84-(a\times 30+b)$	$[84-(a\times 30+b)]^2$
总计	408	1016	$408a+14b$	$1016-(408a+14b)$	$\rightarrow S_e$
平均	29.1	72.6	$29.1a+b$ $=\bar{x}a+b$	$72.6-(29.1a+b)$ $=\bar{y}-(\bar{x}a+b)$	$\frac{S_e}{14}$
	↓ $\bar{x}$	↓ $\bar{y}$			

$$S_e=[77-(a\times 29+b)]^2+\cdots+[84-(a\times 30+b)]^2$$

我们将上表中的 $(y-\hat{y})^2$ 值逐列相加，也就是将 e^2 相加，将其称为“残差平方和”，一般用 S_e 表示。



步骤 3　对残差平方和 S_e 关于 a 和 b 求微分，并使其为 0。

■ 对a微分

$$\frac{dS_e}{da}=2\,[77-(29a+b)]\times(-29)+\cdots+2\,[84-(30a+b)]\times(-30)=0\cdots\cdots①$$

■ 对b微分

$$\frac{dS_e}{db}=2\,[77-(29a+b)]\times(-1)+\cdots+2\,[84-(30a+b)]\times(-1)=0\cdots\cdots②$$

步骤 4　对步骤 3 的①和②进行整理

■ 整理①式

$$2[77-(29a+b)]\times(-29)+\cdots+2[84-(30a+b)]\times(-30)=0$$

$$[77-(29a+b)]\times(-29)+\cdots+[84-(30a+b)]\times(-30)=0$$ 两边同时乘以 $\frac{1}{2}$。

$$29[(29a+b)-77]+\cdots+30[(30a+b)-84]=0$$ 由上式变形可得，请仔细比较。

$$(29\times29a+29\times b-29\times77)+\cdots+(30\times30a+30\times b-30\times84)=0$$

$$(29^2+\cdots+30^2)a+(29+\cdots+30)b-(29\times77+\cdots+30\times84)=0\cdots\cdots③$$

■ 整理②式

$$2[77-(29a+b)]\times(-1)+\cdots+2[84-(30a+b)]\times(-1)=0$$

$$[77-(29a+b)]\times(-1)+\cdots+[84-(30a+b)]\times(-1)=0$$ 两边同时乘以 $\frac{1}{2}$。

$$[(29a+b)-77]+\cdots+[(30a+b)-84]=0$$ 由上式子变形可得，请仔细比较。

$$(29+\cdots+30)a+\underbrace{b+\cdots+b}_{14}-(77+\cdots+84)=0$$

$$(29+\cdots+30)a+14b-(77+\cdots+84)=0$$

$$14b=(77+\cdots+84)-(29+\cdots+30)a$$ 移项得到。

$$b=\frac{77+\cdots+84}{14}-\frac{29+\cdots+30}{14}a\cdots\cdots④$$ 左边只保留b。

$$b=\bar{y}-\bar{x}a\cdots\cdots⑤$$ 请仔细观察上式和步骤2中的表格。

步骤 5 将步骤 4 中的④代入步骤 4 中的③。

$$(29^2+\cdots+30^2)a+(29+\cdots+30)\left(\frac{77+\cdots+84}{14}-\frac{29+\cdots+30}{14}a\right)-(29\times 77+\cdots+30\times 84)=0$$

$$(29^2+\cdots+30)^2a+\frac{(29+\cdots+30)(77+\cdots+84)}{14}-\frac{(29+\cdots+30)^2}{14}a-(29\times 77+\cdots+30\times 84)=0$$

$$\left[(29^2+\cdots+30^2)-\frac{(29+\cdots+30)^2}{14}\right]a+\frac{(29+\cdots+30)(77+\cdots+84)}{14}-(29\times 77+\cdots+30\times 84)=0$$

提取公因式a。

$$\left[(29^2+\cdots+30)^2-\frac{(29+\cdots+30)^2}{14}\right]a=(29\times 77+\cdots+30\times 84)-\frac{(29+\cdots+30)(77+\cdots+84)}{14}$$

移项得到。

整理左边

$$(29^2+\cdots+30^2)-\frac{(29+\cdots+30)^2}{14}$$

$$=(29^2+\cdots+30^2)-2\times\frac{(29+\cdots+30)^2}{14}+\frac{(29+\cdots+30)^2}{14}$$

由上式变形可得，请仔细比较。

$$=(29^2+\cdots+30^2)-2\times(29+\cdots+30)\times\frac{29+\cdots+30}{14}+\left(\frac{29+\cdots+30}{14}\right)^2\times 14$$

$$=(29^2+\cdots+30^2)-2\times(29+\cdots+30)\times\bar{x}+(\bar{x})^2\times 14$$

$\bar{x}=\frac{29+\cdots+30}{14}$

$$=(29^2+\cdots+30^2)-2\times(29+\cdots+30)\times\bar{x}+\underbrace{(\bar{x})^2+\cdots+(\bar{x})^2}_{14}$$

$$=[29^2-2\times 29\times\bar{x}+(\bar{x})^2]+\cdots+[30^2-2\times 30\times\bar{x}+(\bar{x})^2]$$

$$=(29-\bar{x})^2+\cdots+(30-\bar{x})^2$$

$$=S_{xx}$$

整理右边

$$(29\times 77+\cdots+30\times 84)-\frac{(29+\cdots+30)(77+\cdots+84)}{14}$$

$$=(29\times 77+\cdots+30\times 84)-\frac{29+\cdots+30}{14}\times\frac{77+\cdots+84}{14}\times 14$$

$$=(29\times 77+\cdots+30\times 84)-\bar{x}\times\bar{y}\times 14$$

$$=(29\times 77+\cdots+30\times 84)-\bar{x}\times\bar{y}\times 14-\bar{x}\times\bar{y}\times 14+\bar{x}\times\bar{y}\times 14$$

由上式变形可得，请仔细比较。

$$=(29\times 77+\cdots+30\times 84)-\frac{29+\cdots+30}{14}\times\bar{y}\times 14-\bar{x}\times\frac{77+\cdots+84}{14}\times 14+\bar{x}\times\bar{y}\times 14$$

$$=(29\times 77+\cdots+30\times 84)-(29+\cdots+30)\bar{y}-\bar{x}(77+\cdots+84)+\bar{x}\times\bar{y}\times 14$$

$$=(29\times 77+\cdots+30\times 84)-(29+\cdots+30)\bar{y}-(77+\cdots+84)\bar{x}+\underbrace{\bar{x}\times\bar{y}+\cdots+\bar{x}\times\bar{y}}_{14}$$

$$=(29\times 77-29\bar{y}-77\bar{x}+\bar{x}\times\bar{y})+\cdots+(30\times 84-30\bar{y}-84\bar{x}+\bar{x}\times\bar{y})$$

$$=(29-\bar{x})(77-\bar{y})+\cdots+(30-\bar{x})(84-\bar{y})$$

$$=S_{xy}$$

$$S_{xx}a=S_{xy}$$

$$a=\frac{S_{xy}}{S_{xx}}\quad\cdots\cdots ⑥$$

左边只保留a。

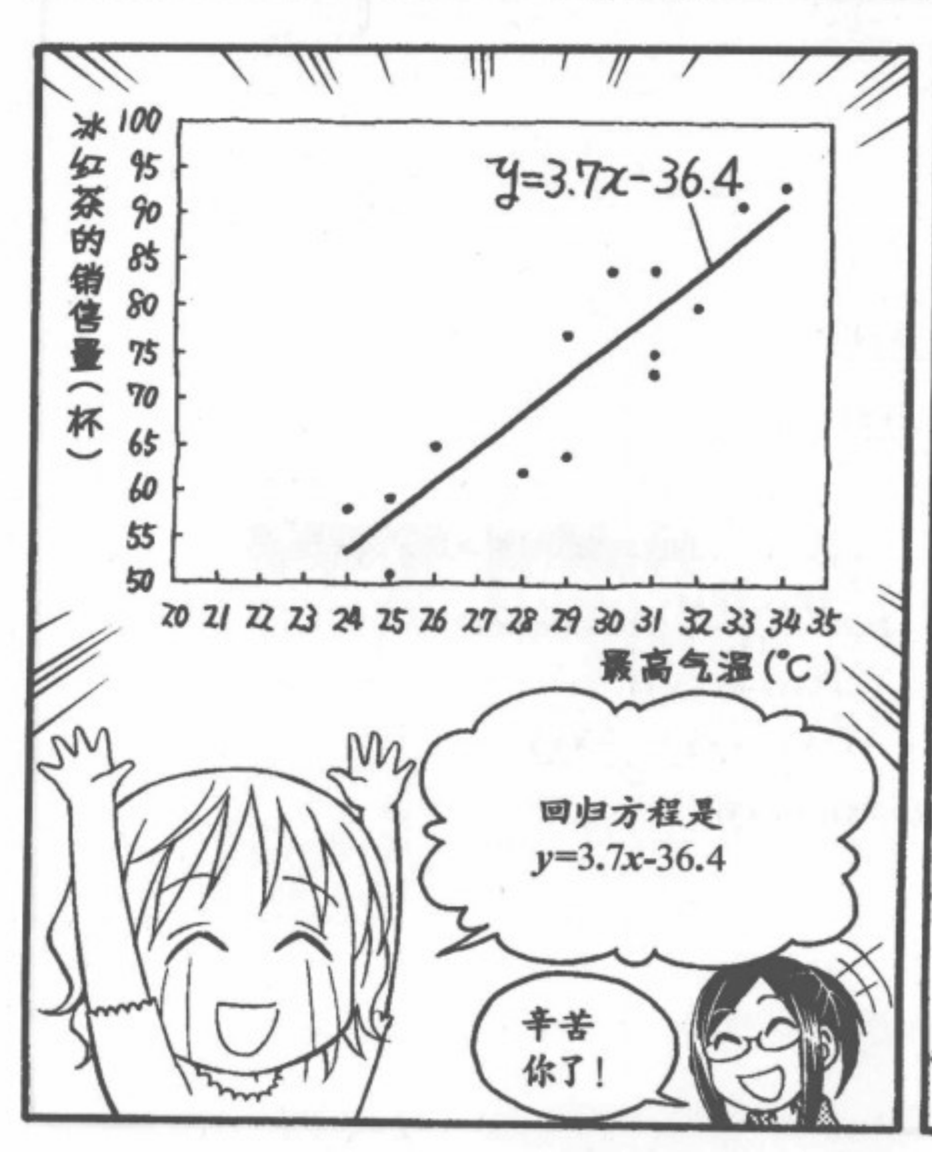
步骤 6 求解回归方程：

根据步骤 5 中的⑥可知，$a=\dfrac{S_{xy}}{S_{xx}}$。根据步骤 4 中的⑤可知，$b=\bar{y}-\bar{x}a$。因此，根据步骤 1 可知

$$\begin{cases} a=\dfrac{S_{xy}}{S_{xx}}=\dfrac{484.9}{129.7}=3.7 \\ b=\bar{y}-\bar{x}a=72.6-29.1\times 3.7=-36.4 \end{cases}$$

所以，回归方程为

$$y=3.7x-36.4$$

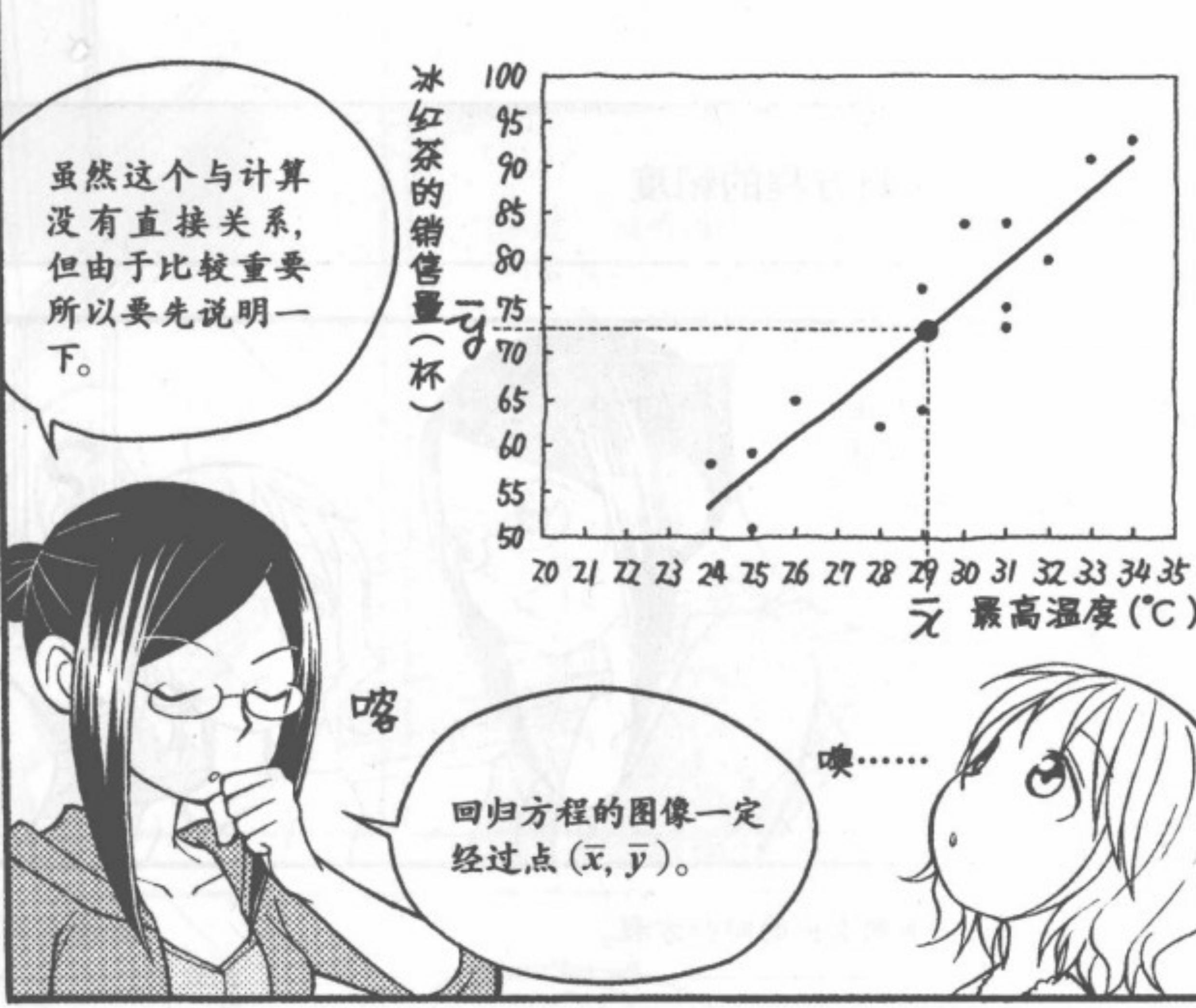


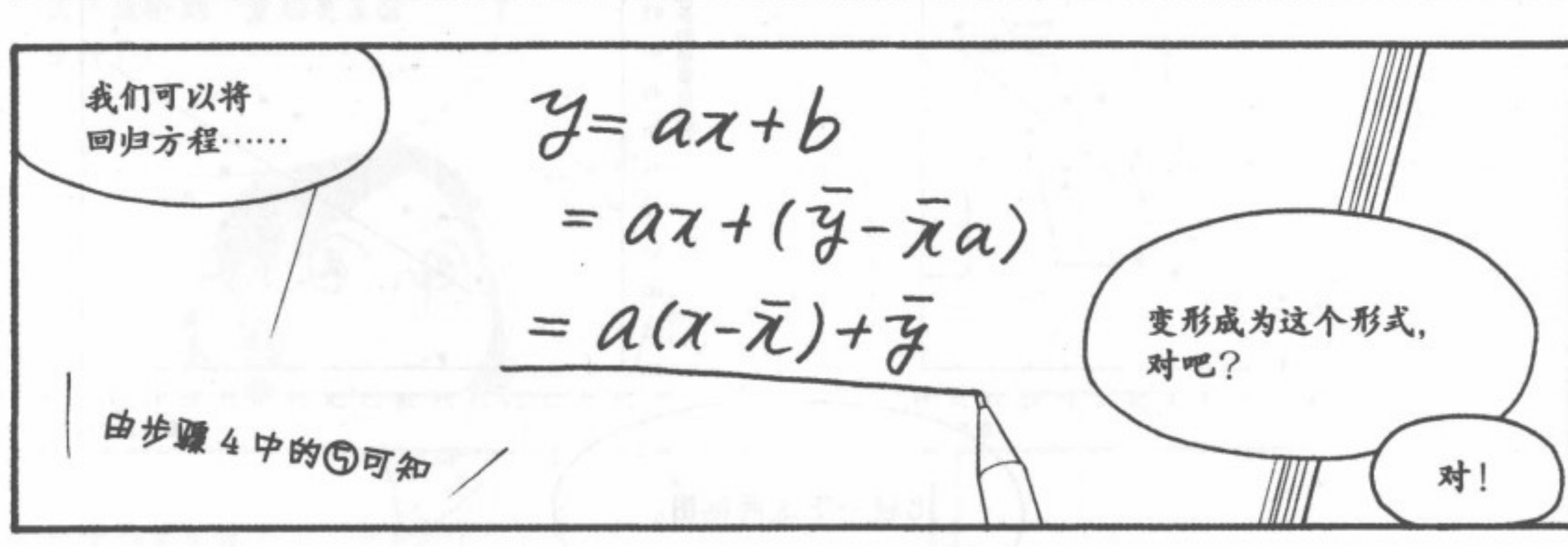
不仅仅限于这个例子，凡是求回归方程的 a 和 b 的值，都可以做如下计算。

$$\begin{cases} a=\dfrac{x\text{ 和 }y\text{ 的离差积和}}{x\text{ 的离差平方和}}=\dfrac{S_{xy}}{S_{xx}} \\ b=\bar{y}-x\bar{a} \end{cases}$$

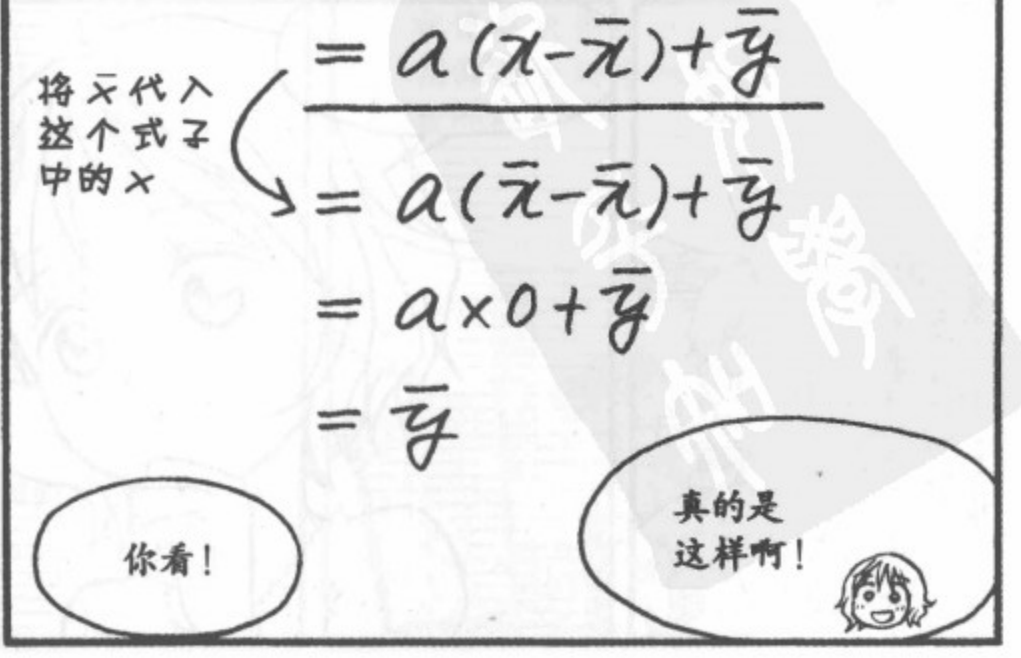
要记住啊！


啊，美羽，最高气温和冰红茶的销售量的平均值分别是多少？
嗯……
29.1和72.6。

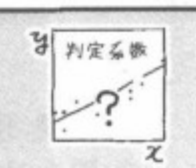

虽然这个与计算没有直接关系，但由于比较重要所以要先说明一下。
冰红茶的销售量（杯）
最高温度（℃）
喀
回归方程的图像一定经过点 $(\bar{x}, \bar{y})$。
噢……

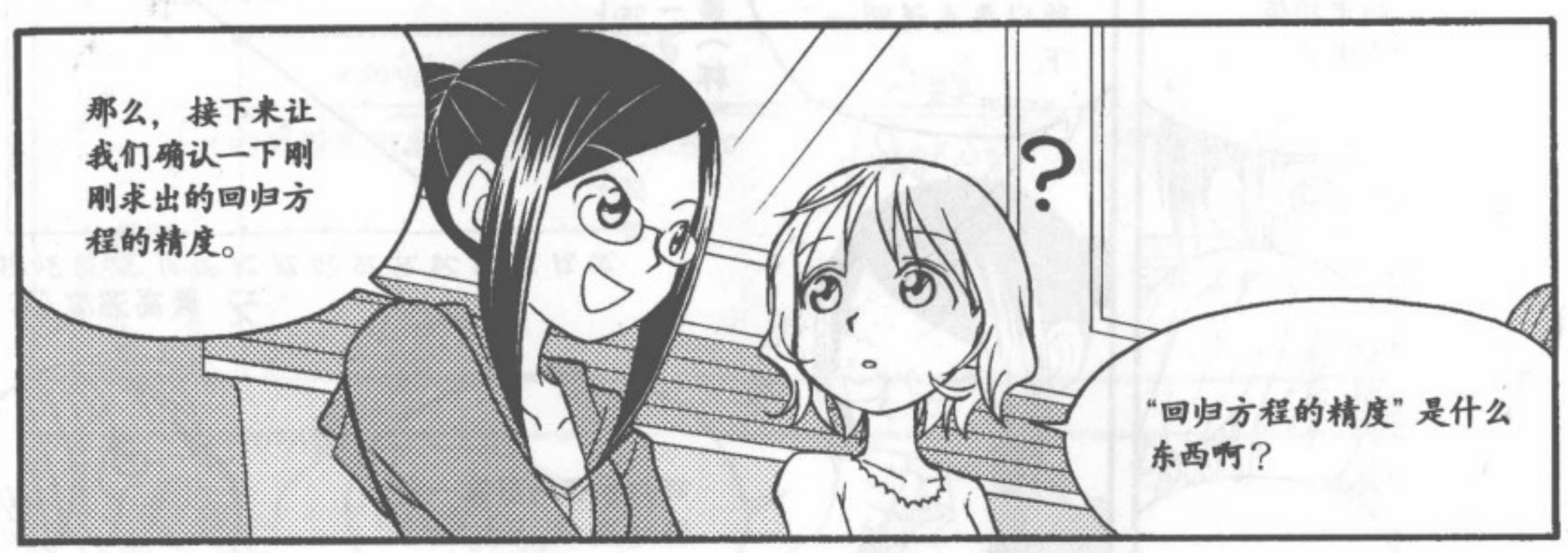

我们可以将回归方程……
$y = ax + b$
$= ax + (\bar{y} - \bar{x}a)$
$= a(x - \bar{x}) + \bar{y}$
由步骤4中的⑤可知
变形成为这个形式，对吧？
对！


将 $\bar{x}$ 代入这个式子中的 x。


$= a(x - \bar{x}) + \bar{y}$
将 $\bar{x}$ 代入这个式子中的 x
$= a(\bar{x} - \bar{x}) + \bar{y}$
$= a \times 0 + \bar{y}$
$= \bar{y}$
你看！
真的是这样啊！

③ 确认回归方程的精度





刚刚求出的回归方程。

冰红茶的销售量(杯)

$y = 3.7x - 36.4$

最高温度(℃)

根据虚构的数据求出的回归方程。

冰红茶的销售量(杯)

最高温度(℃)

比较一下这两幅图。




像这样，我们把左图中点和回归方程吻合得好的情况，就叫做“回归方程的精度较高”或者“拟合度较好”。
哦！


不过，以看图的方式来判断精度，毕竟还是有些主观，你说对吗？
这个拟合得好一些
这个拟合得不够好
是啊……


这下该轮到“重相关系数”出场了！
通常用“R”来表示！


重相关系数？
回归方程的精度就是用来表示点和回归方程拟合程度的指标。


而重相关系数说起来就像单次测量值 y 和预测值 ŷ 的单相关系数那样。
哦！



$$R=\frac{y\text{和}\hat{y}\text{的离差积和}}{\sqrt{y\text{的离差平方和}\times\hat{y}\text{的离差平方和}}}=\frac{S_{y\hat{y}}}{\sqrt{S_{yy}\times S_{\hat{y}\hat{y}}}}=\frac{1812.3}{\sqrt{2203.4\times1812.3}}=0.9069$$

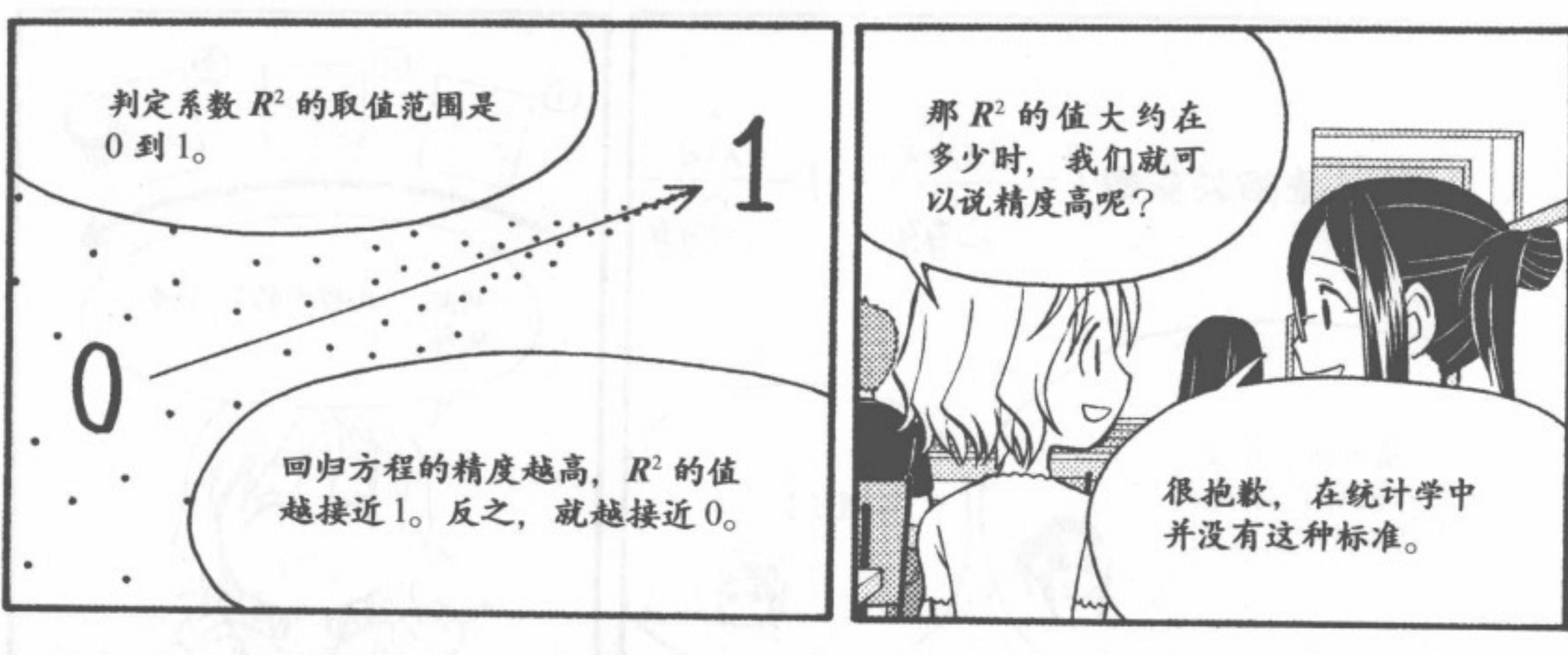


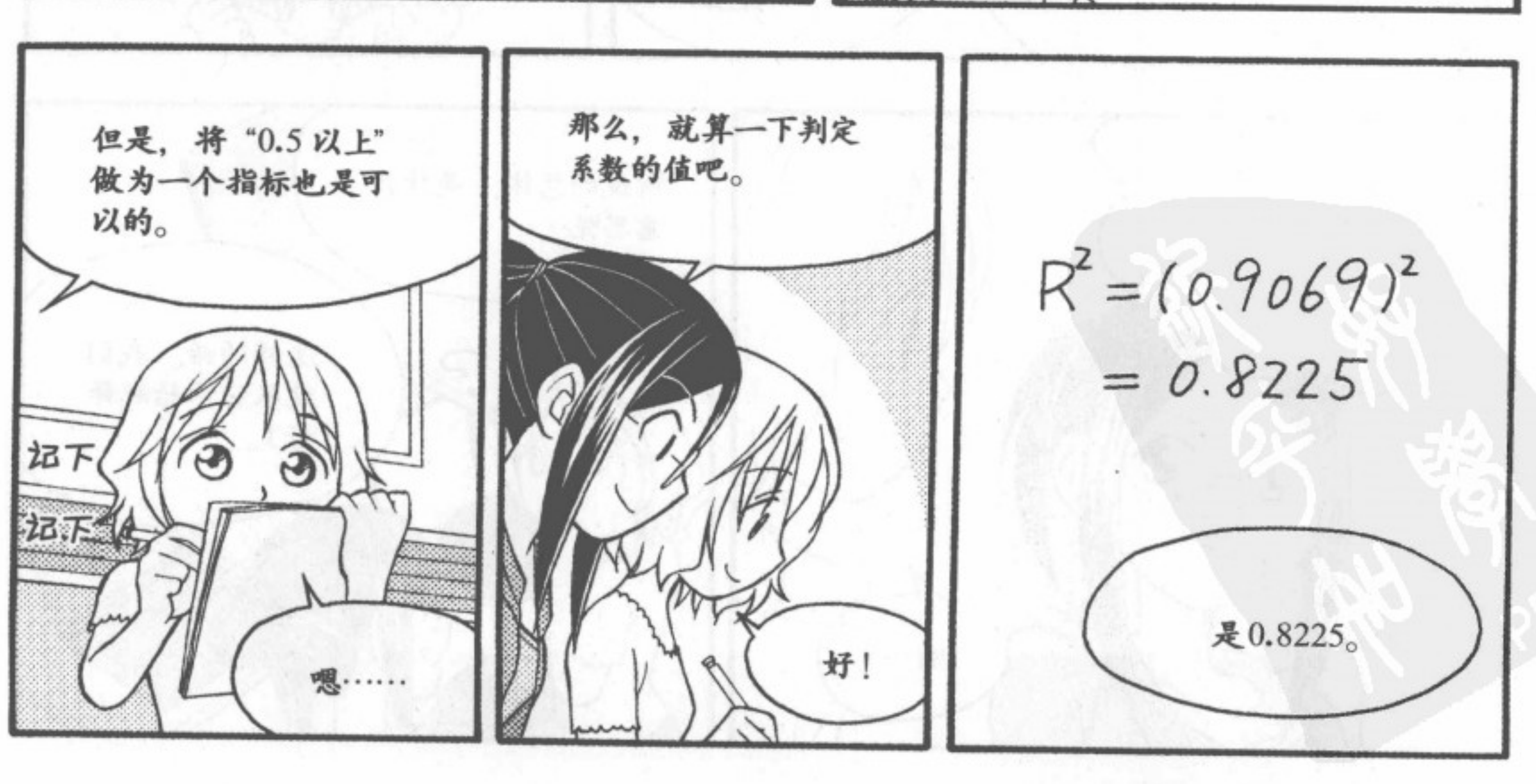
计算过程!

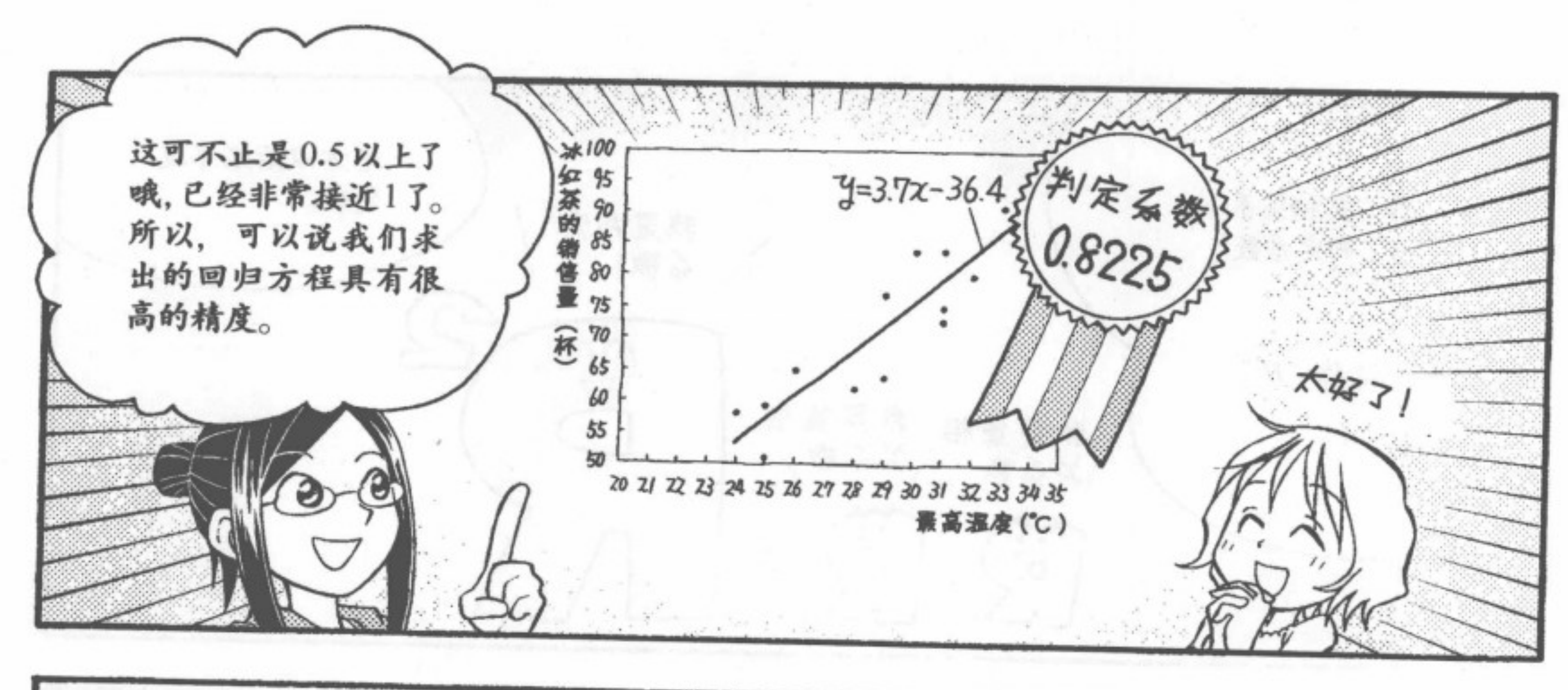
	实测值 y	预测值 $\hat{y}=3.7x-36.4$	$y-\bar{y}$	$\hat{y}-\bar{\hat{y}}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{\hat{y}})^2$	$(y-\bar{y})(\hat{y}-\bar{\hat{y}})$	$(y-\hat{y})^2$
22日 (一)	77	72.0	4.4	−0.5	19.6	0.3	−2.4	24.6
23日 (二)	62	68.3	−10.6	−4.3	111.8	18.2	45.2	39.7
24日 (三)	93	90.7	20.4	18.2	417.3	329.6	370.9	5.2
25日 (四)	84	79.5	11.4	6.9	130.6	48.2	79.3	20.1
26日 (五)	59	57.1	−13.6	−15.5	184.2	239.8	210.2	3.7
27日 (六)	64	72.0	−8.6	−0.5	73.5	0.3	4.6	64.6
28日 (日)	80	83.3	7.4	10.7	55.2	114.1	79.3	10.6
29日 (一)	75	79.5	2.4	6.9	5.9	48.2	16.9	20.4
30日 (二)	58	53.3	−14.6	−19.2	212.3	369.5	280.1	21.6
31日 (三)	91	87.0	18.4	14.4	339.6	207.9	265.7	16.1
1日 (四)	51	57.1	−21.6	−15.5	465.3	239.8	334.0	37.0
2日 (五)	73	79.5	0.4	6.9	0.2	48.2	3.0	42.4
3日 (六)	65	60.8	−7.6	−11.7	57.3	138.0	88.9	17.4
4日 (日)	84	75.8	11.4	3.2	130.6	10.3	36.6	67.6
总 计	1016	1016	0	0	**2203.4**	**1812.3**	**1812.3**	391.1
平 均	72.6	72.6			↓ S_{yy}	↓ $S_{\hat{y}\hat{y}}$	↓ $S_{y\hat{y}}$	↓ S_e
	↓ $\bar{y}$	↓ $\bar{\hat{y}}$						

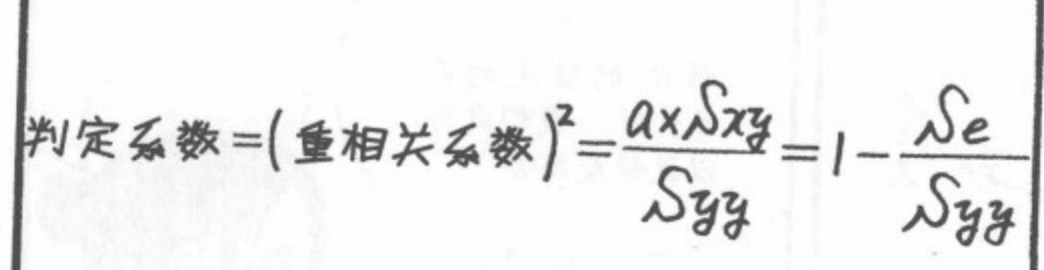
重相关系数 R 的计算虽然与 S_e 没有关系，但是对之后的运算很重要，所以要先求出来。


我们将（重相关系）^{2}数称为“判定系数”。
通常，记作“R^2”。
我是重相关系数
我是重相关系数
我是判定系数！
平方以后名字就变啦。


判定系数 R^2 的取值范围是0到1。
1
0
回归方程的精度越高，R^2 的值越接近1。反之，就越接近0。
那 R^2 的值大约在多少时，我们就可以说精度高呢？
很抱歉，在统计学中并没有这种标准。


但是，将“0.5以上”做为一个指标也是可以的。
记下
记下
嗯……
那么，就算一下判定系数的值吧。
好！
$R^2 = (0.9069)^2 = 0.8225$
是0.8225。



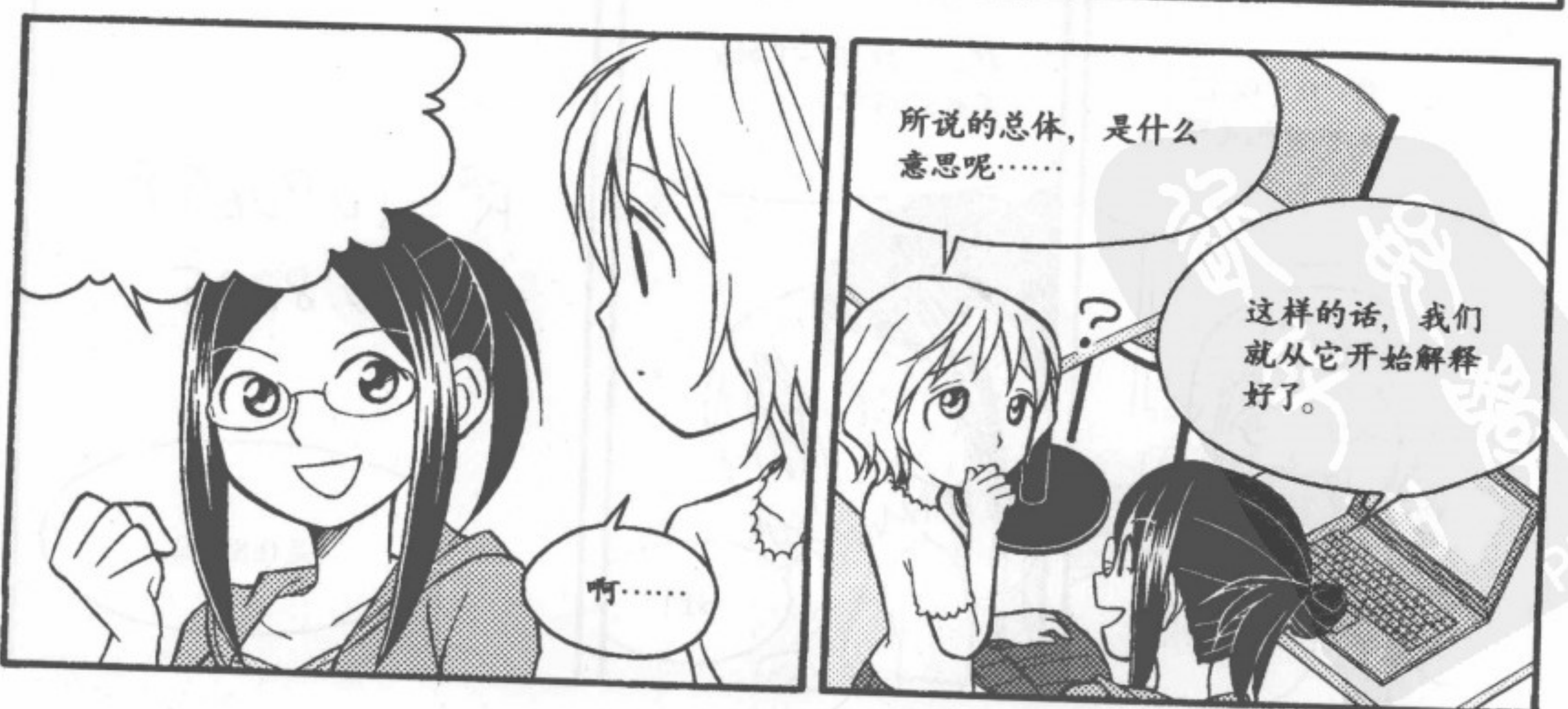




为了节省时间，这里就不作证明了。不过，这样的关系可是成立的哟！



知道了！

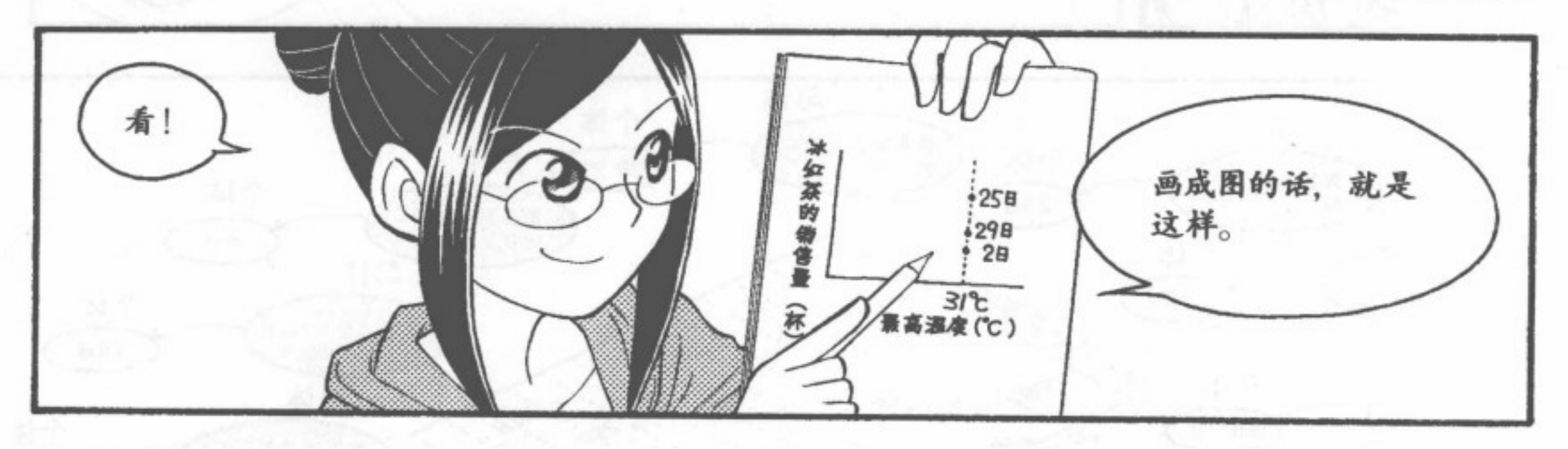




	最高温度 ℃	冰红茶的销售量（杯）
22日（一）	29	77
23日（二）	28	62
24日（三）	34	93
25日（四）	31	84
26日（五）	25	59
27日（六）	29	64
28日（日）	32	80
29日（一）	31	75
30日（二）	24	58
31日（三）	33	91
1日（四）	25	51
2日（五）	31	73
3日（六）	26	65
4日（日）	30	84

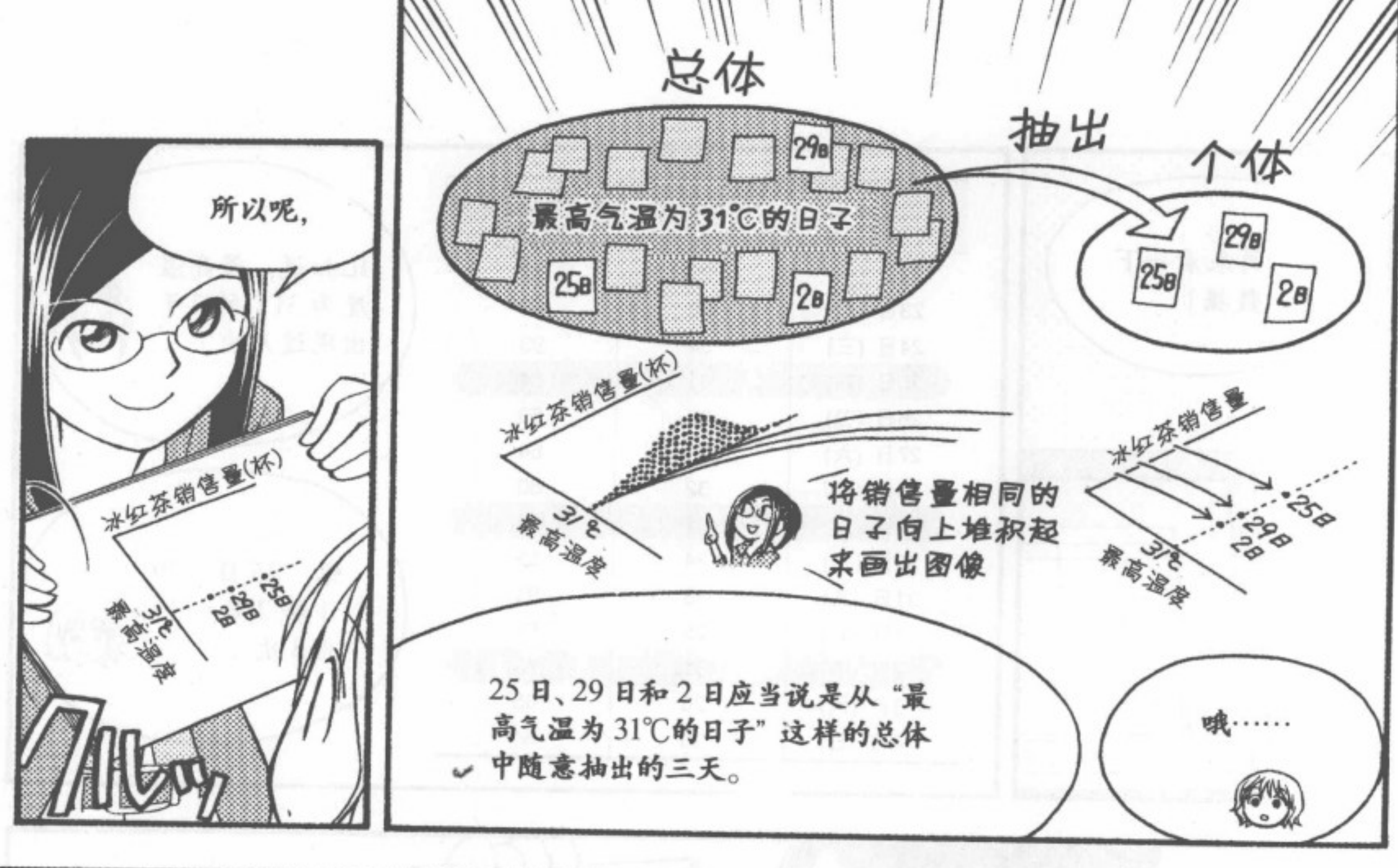












总体
最高温度为28℃的日子
个体
23日

总体
最高温度为29℃的日子
个体
22日 27日

总体
最高温度为30℃的日子
个体
4日

总体
最高温度为26℃的日子
个体
3日

总体
最高温度为32℃的日子
个体
28日

总体
最高温度为25℃的日子
个体
26日 1日

总体
最高温度为33℃的日子
个体
31日

总体
最高温度为24℃的日子
个体
30日

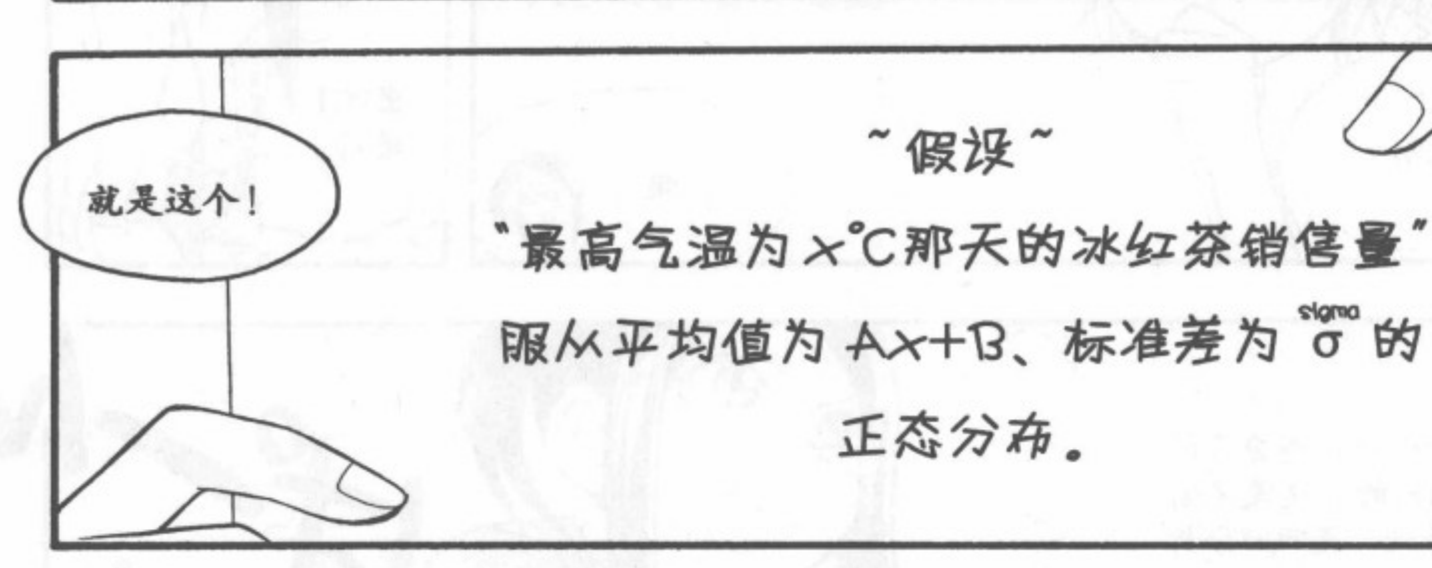
总体
最高温度为34℃的日子
个体
24日

冰红茶销售量(杯)
30日 1日 26日 3日 23日 27日 22日 4日 2日 29日 25日 28日 31日 24日
24 25 26 27 28 29 30 31 32 33 34 35
最高气温(℃)

其他的日子也这样做吧！
您这么一说……

我明白所谓的总体是怎么一回事了。
好样的！那我们继续吧！





符号的含义以后再解释，我们先来看一下这个假设！

冰红茶的销售量(杯)

$Ax+B$

相同的图形

26
28
30
32

最高温度(℃)

？

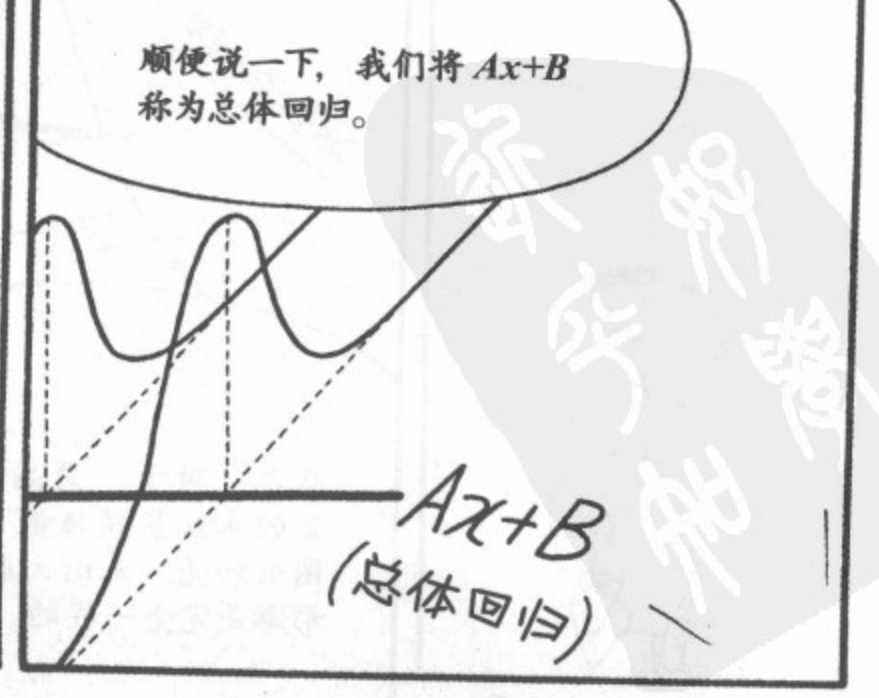
总之，对于“最高气温为x℃那天的冰红茶销售量”的正态分布图形来说，无论x取什么值，图形都是完全一样的。


"完全一样"指的是……
寒冷的日子
炎热的日子
我觉得好像是根据气温的不同而不同啊！
那样的情况有可能吗？
唉！

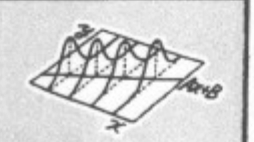

就是那样……一定是那样的……
来吧！美羽。


记住，无论冷热正态分布的图形都是相同的，这是不需要任何解释的，是回归分析的规定！
ぱーん
ざっ


这样说的话，是不是就可以想通了？
是！

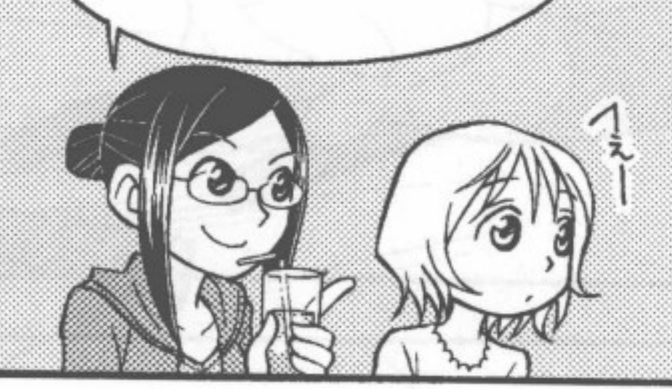

顺便说一下，我们将 Ax+B 称为总体回归。
Ax+B
(总体回归)

④ 进行"回归系数的检验"



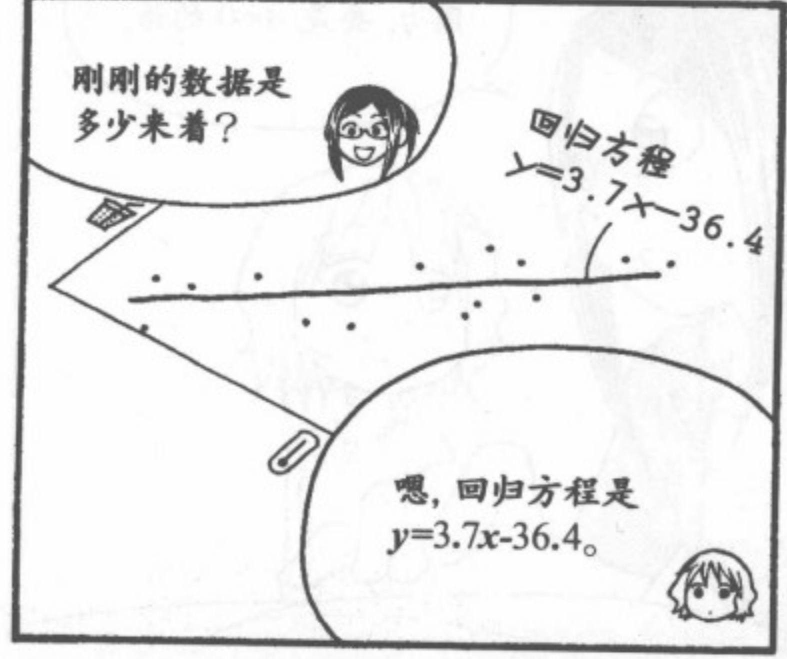


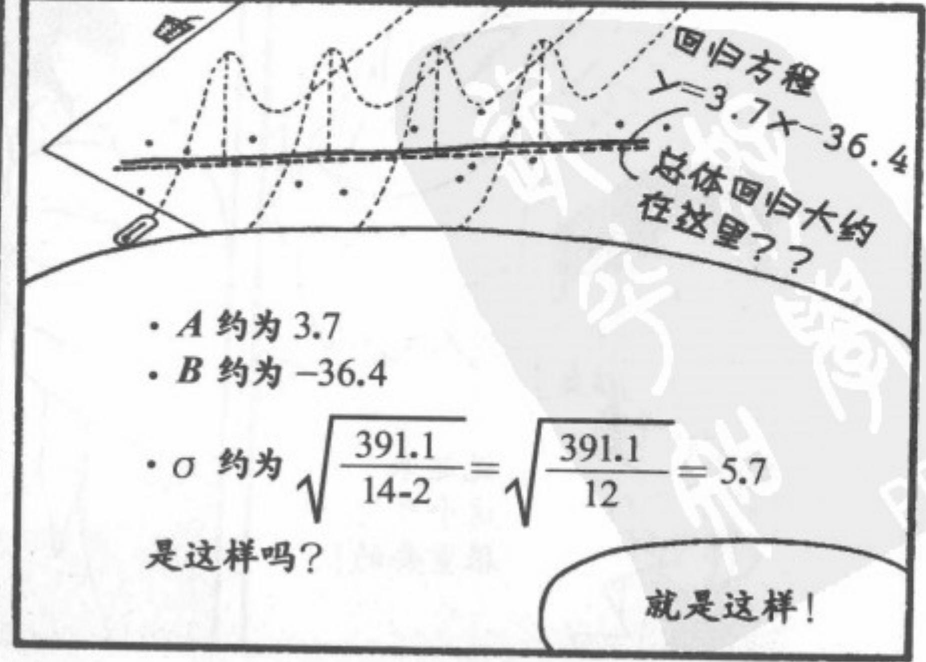




在刚刚求出的回归方程 $y=ax+b$ 中

- A 约为 a
- B 约为 b
- σ 约为 $\sqrt{\dfrac{S_e}{\text{个体个数}-2}}$





- A 约为 3.7
- B 约为 −36.4
- σ 约为 $\sqrt{\dfrac{391.1}{14-2}}=\sqrt{\dfrac{391.1}{12}}\doteq 5.7$

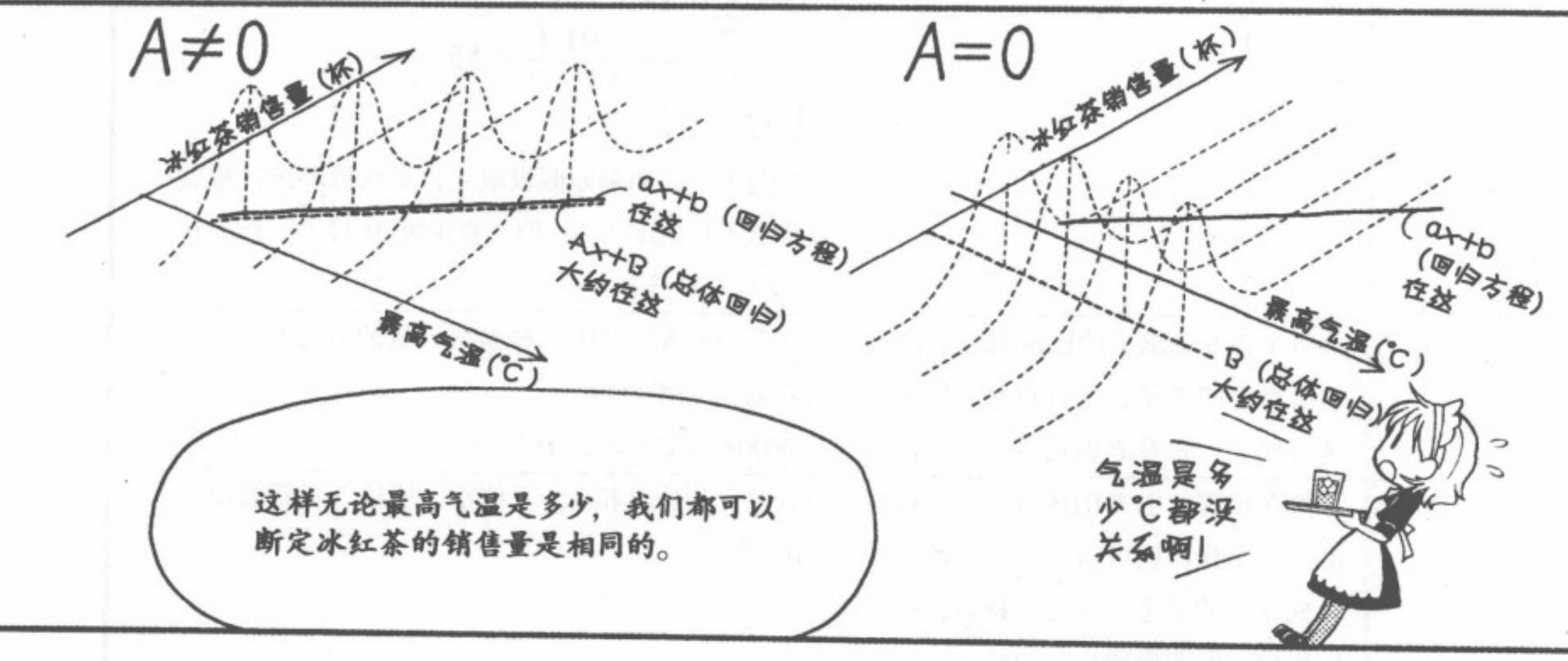
是这样吗？

就是这样！

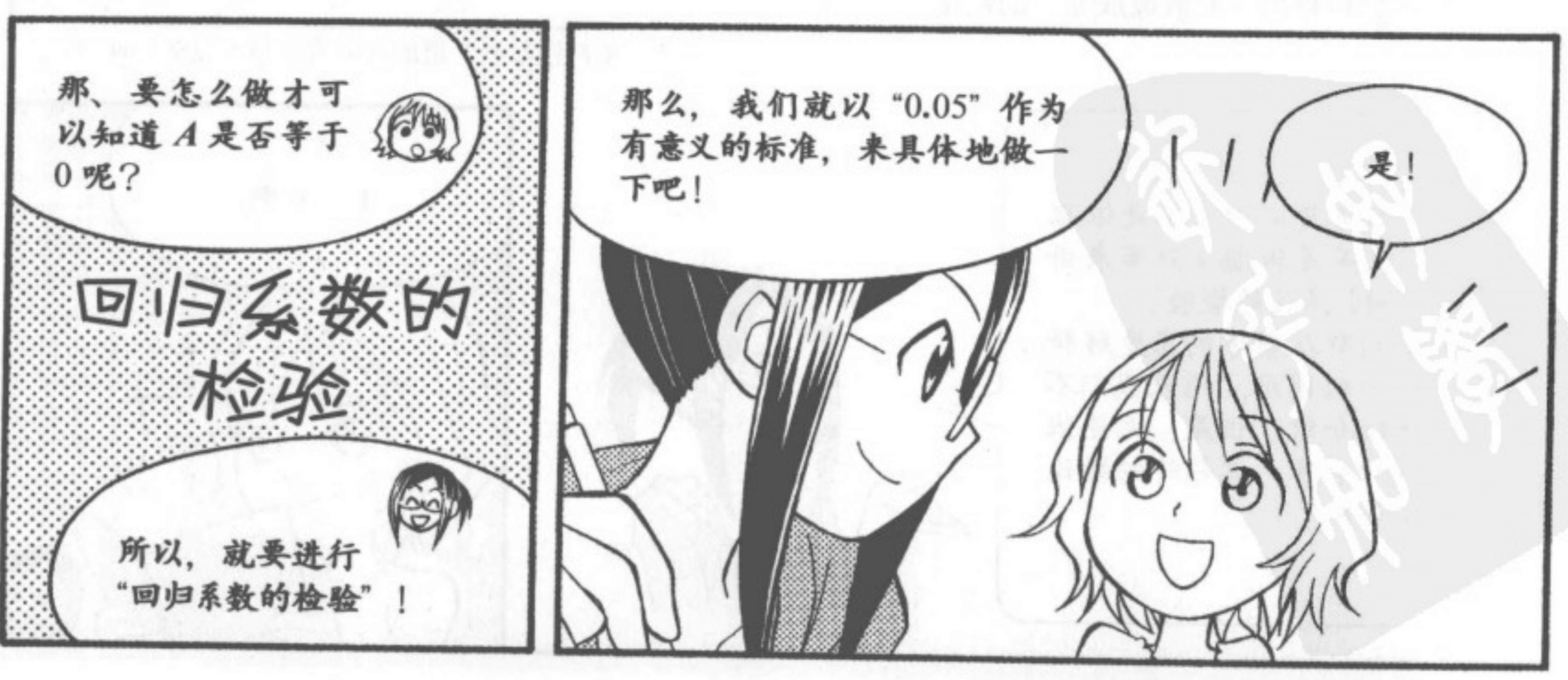

"约为……"说起来
不是很模糊吗？
那是因为
运用统计学的手段，
无法严格地计算出 A、B 和
σ 的值．
还是不明白，是吗？
但是……
要记住
"A 的值是
无论如何
都不能等于 0 的"！
……
说起来
这个可是
很重要的！
因为，要是 A=0 的话，



就会发生这样的情况。
~假设~
“最高气温为x℃那天的冰红茶销售量”服从平均值为B、标准差为σ的正态分布。
啊！



A≠0
冰红茶销售量（杯）
ax+b在这（回归方程）
Ax+B大约在这（总体回归）
最高气温（℃）
A=0
冰红茶销售量（杯）
ax+b在这（回归方程）
B大约在这（总体回归）
最高气温（℃）
这样无论最高气温是多少，我们都可以断定冰红茶的销售量是相同的。
气温是多少℃都没关系啊！



那、要怎么做才可以知道A是否等于0呢？
回归系数的检验
所以，就要进行“回归系数的检验”！
那么，我们就以“0.05”作为有意义的标准，来具体地做一下吧！
是！

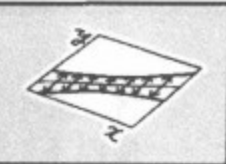
步骤 1	定义总体。	以“最高气温为 x℃的日子”作为总体。
步骤 2	建立原假设和备择假设。	原假设为“$A=0$”。 备择假设为“$A\neq 0$”。
步骤 3	选择所要进行的“检验”类型。	进行“回归系数的检验”。
步骤 4	设定有意义的标准。	以 0.05 为有意义的标准。
步骤 5	通过样本数据求出检验统计量的值。	下面进行“回归系数的检验”。 所谓“回归系数的检验”的检验统计量的值为 $\dfrac{a^2}{\left(\dfrac{1}{S_{xx}}\right)} \div \dfrac{S_e}{\text{个体个数}-2}$ 所以在本例题中，检验统计量的值为 $\dfrac{3.7^2}{\left(\dfrac{1}{129.7}\right)} \div \dfrac{391.1}{14-2} = 55.6$ 在本例题中，如果原假设成立，那么检验统计量就服从第 1 自由度为 1、第 2 自由度为 12（= 个体个数 −2）的 F 分布。
步骤 6	再将步骤 5 中求出的检验统计量的值所对应的 P 值，与有意义的标准进行比较，看看 P 值是否比其小。	有意义的标准是 0.05。检验统计量的值为 55.6，所以 P 值为 0.000008。 0.000008<0.05，所以 P 值小。
步骤 7	如果在步骤 6 中 P 值比有意义的标准小，则我们就可以得出“备择假设成立”的结论。反之，我们就可以得出“原假设成立”的结论。	与有意义的标准相比，P 值小。所以，备择假设“$A\neq 0$”成立。

※F 分布中 P 值的求解方法请参见第 204 页。

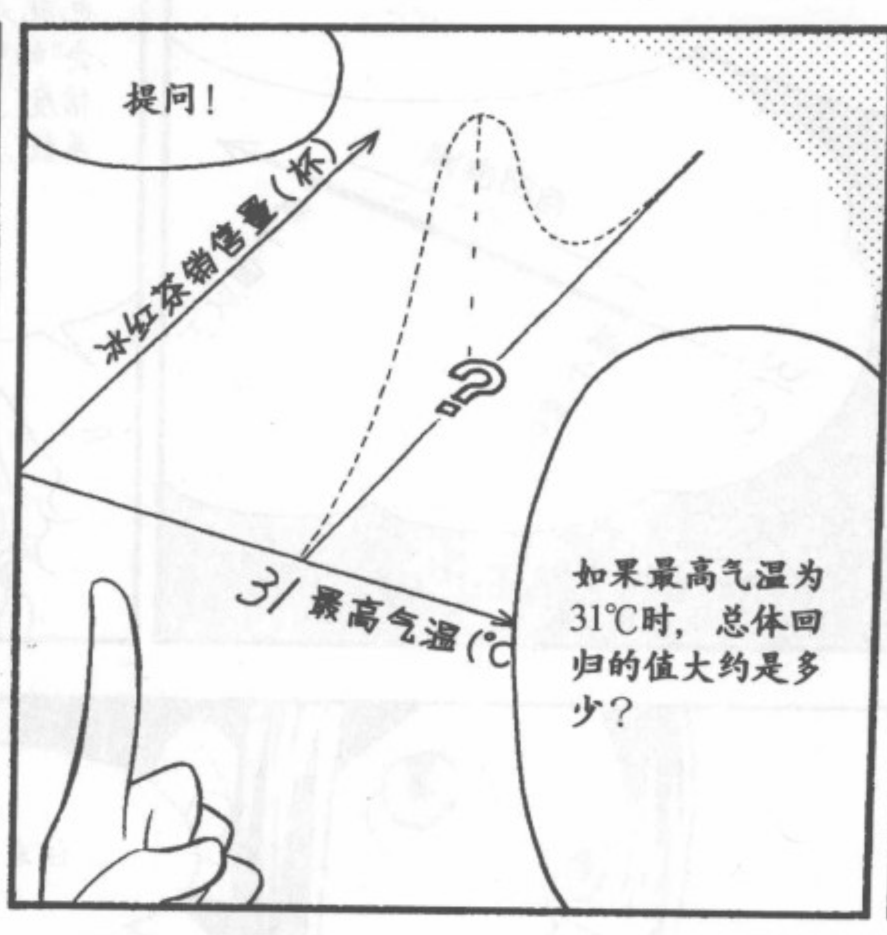




⑤ 总体回归 $Ax+B$ 的估计









但是有趣的是，

在统计学中，"总体回归 $Ax+B$'一定会'在'某个值以上、某个值以下'的区间中"，这种说法要理解哟！

冰红茶销售量（杯）

某个值以下

（回归方程）

某个值以上

Ax31+B 应该位于这个区间中的某个部分！

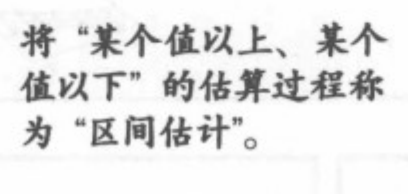
15

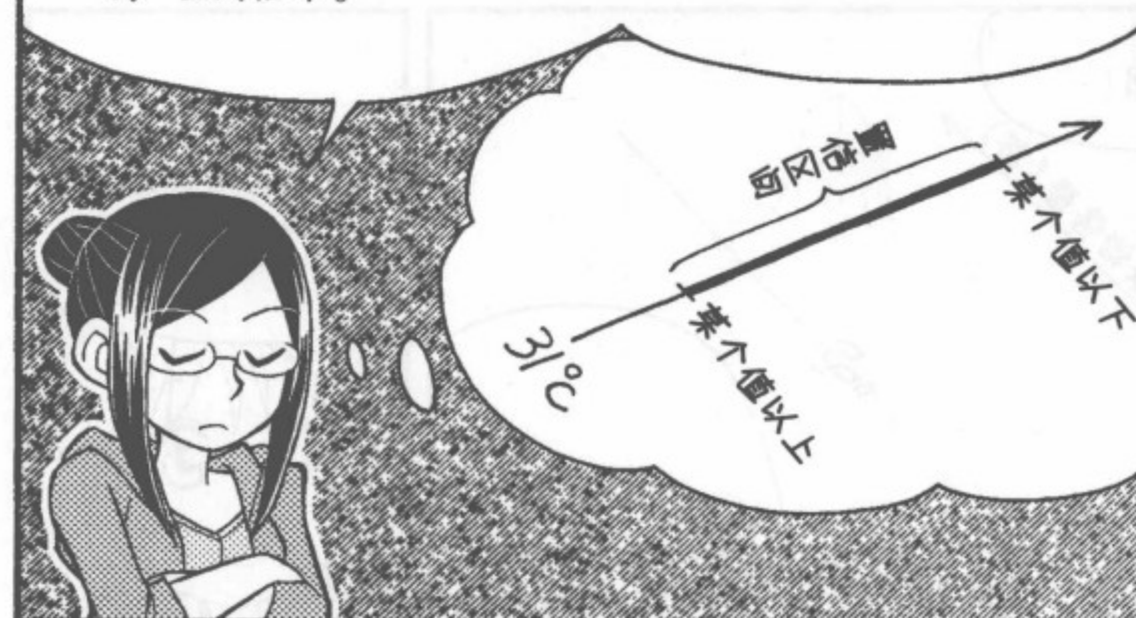
$\bar{x}$

31

最高温度（℃）

不同的x值对应着不同的宽度啊！









置信度并不是在求出置信区间后判断出来的。

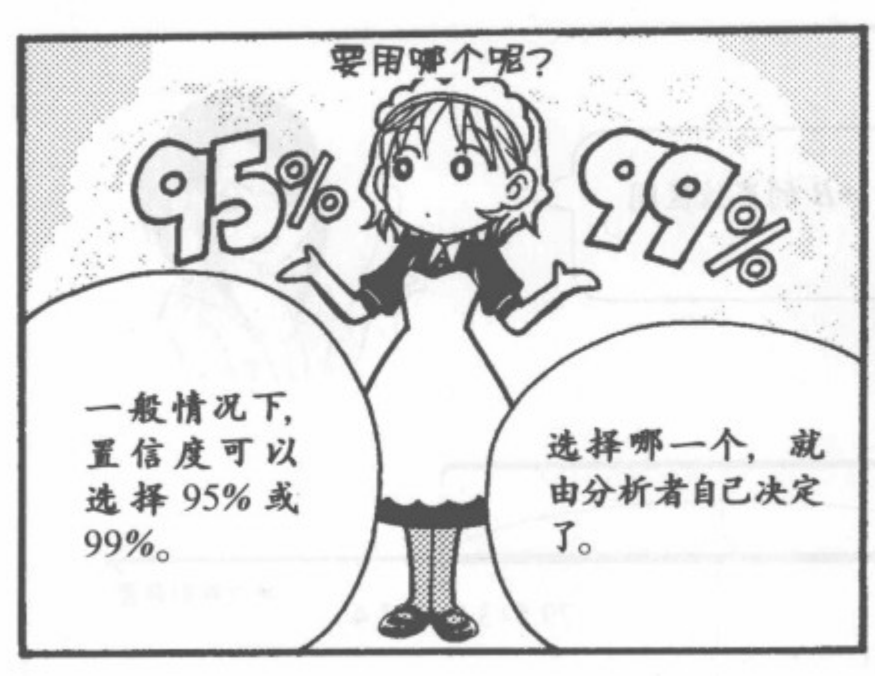
实际上，它是在求解置信区间之前，由分析者自己"决定"的。



在刚刚的31℃的例子中，实际上并不是说总体回归 $A \times 31 + B$ "一定会"位于"某个值以上、某个值以下"，

而应当说，当置信度为〇〇%时，总体回归 $A \times 31 + B$ 应该位于"某个值以上、某个值以下"的区间内。







虽然说99%的可靠性确实很高，可是……

99%

冰红茶的销售量大概会是0～120杯，应该没错的！

呃……

这个…

店主

95%

冰红茶的销售量大概会是40～80杯，基本不会有错！

嗯……

店主

那样就会导致置信区间变长，也就没有说服力了。

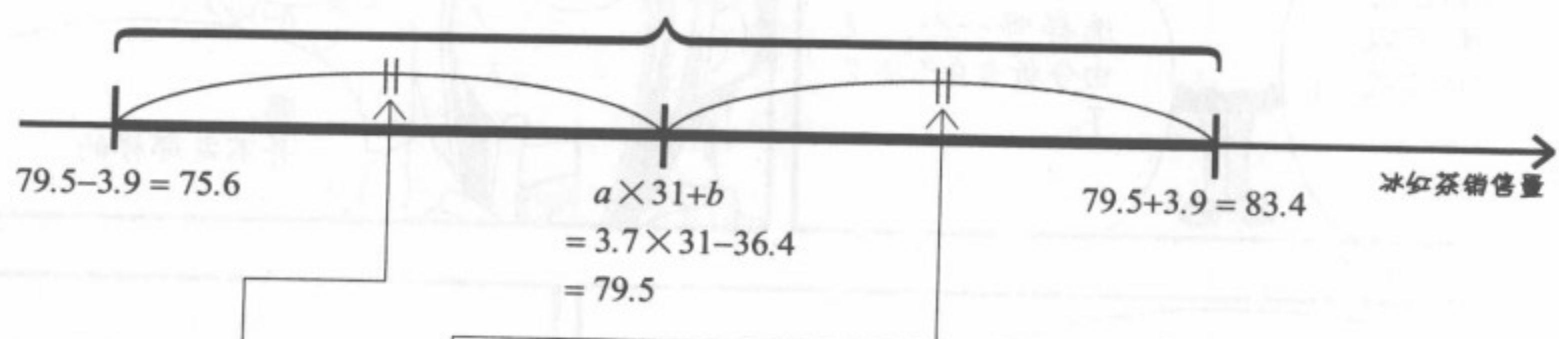
的确如此！

那么，你来求一下，置信度为95%时的总体回归 $A \times 31 + B$ 的置信区间吧！

是！

当置信度为 95% 的时候，总体回归 $A\times31+B$ 的置信区间如下：





这两部分的长度都是

$$\sqrt{F(1，样本个数-2；0.05)\times\left[\frac{1}{样本个数}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]\times\frac{S_e}{样本个数-2}}$$

$$=\sqrt{F(1，14-2；0.05)\times\left[\frac{1}{14}+\frac{(31-29.1)^2}{129.7}\right]\times\frac{391.1}{14-2}}$$

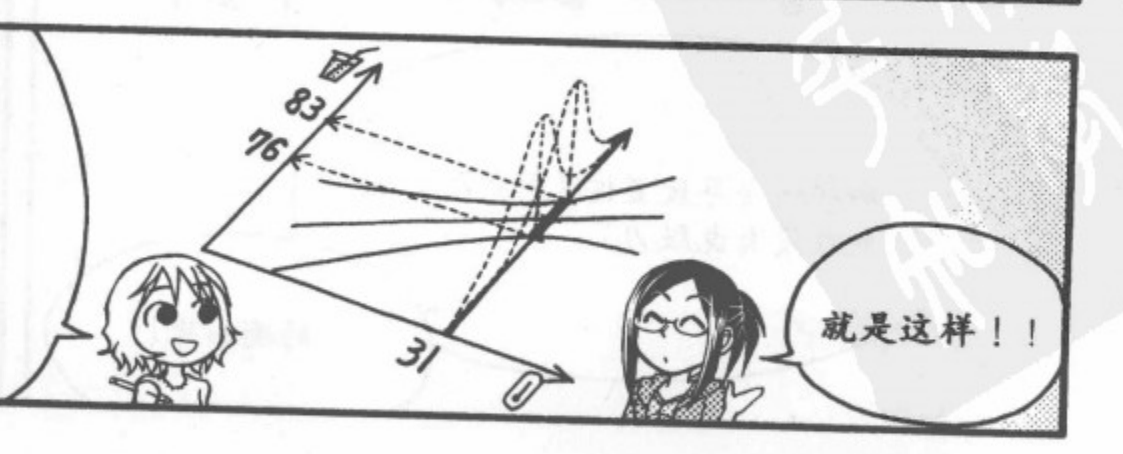
$= 3.9$

当置信度为 99% 的时候，总体回归 $A\times31+B$ 的置信区间只需要将 $F(1，样本个数-2；0.05) = F(1，14-2；0.05) = 4.7$ 的部分变成 $F(1，样本个数-2；0.01) = F(1，14-2；0.01) = 9.3$

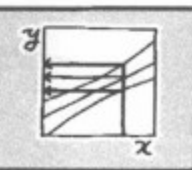


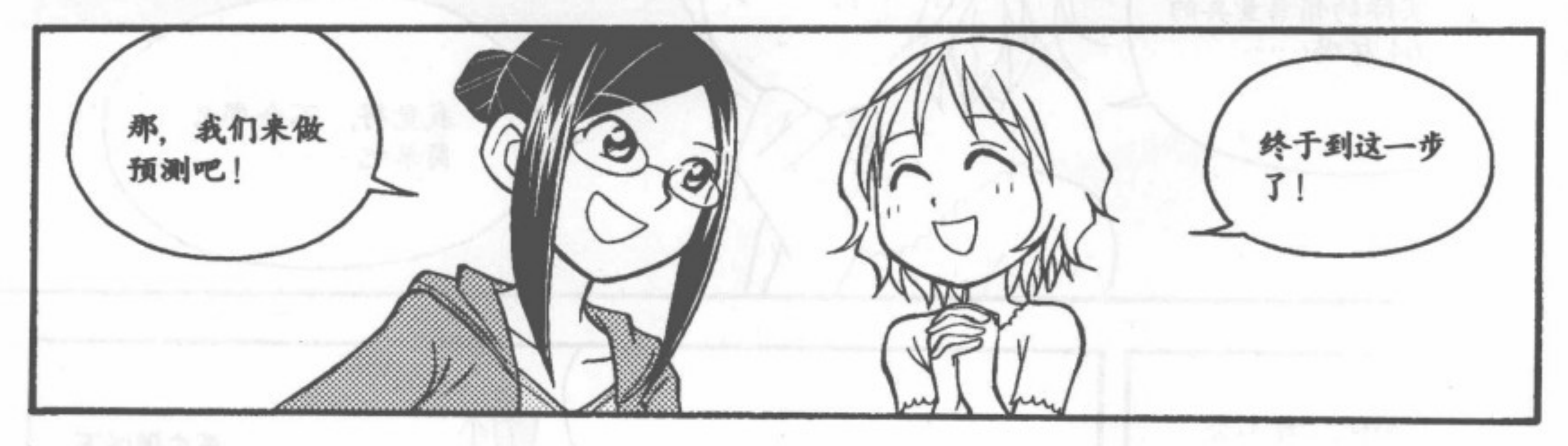
※关于 $F(1,14-2;0.05)=4.7$ 等的介绍，请参见第 54 页。

在置信度为 95% 时，总体回归 $A\times31+B$ 处于 76 杯以上、83 杯以下是这样的吗？



⑥预 测





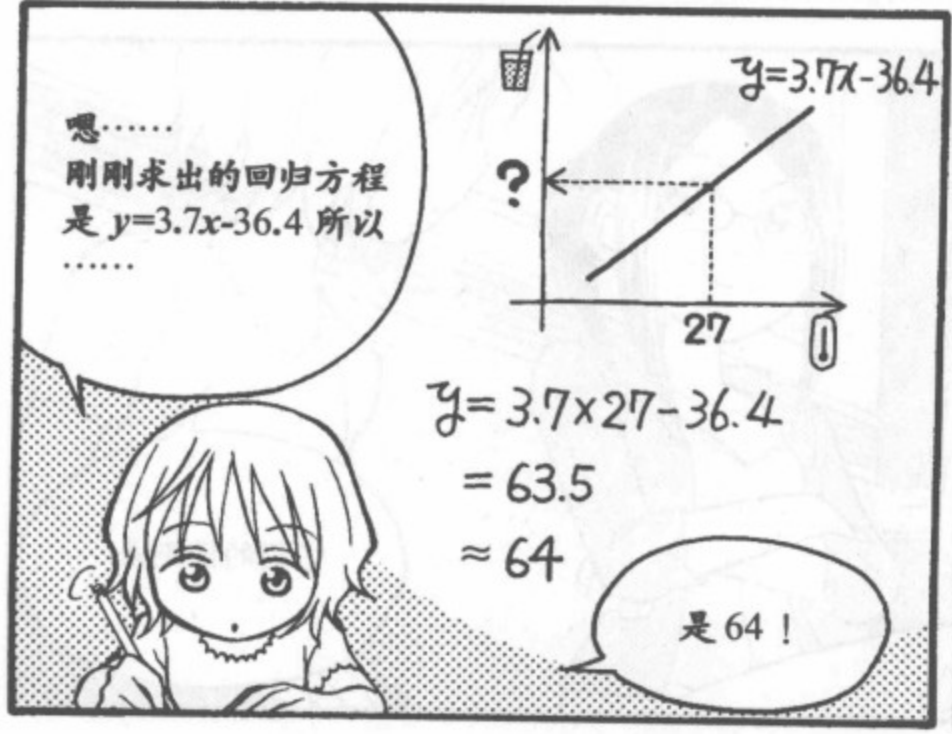
终于到这一步了！
那，我们来做预测吧！



明天天气
晴
最高气温27℃
最低气温20℃
降水概率0%
明天的最高气温是……27℃啊！



那么，明天的冰红茶的销售量会是多少呢？



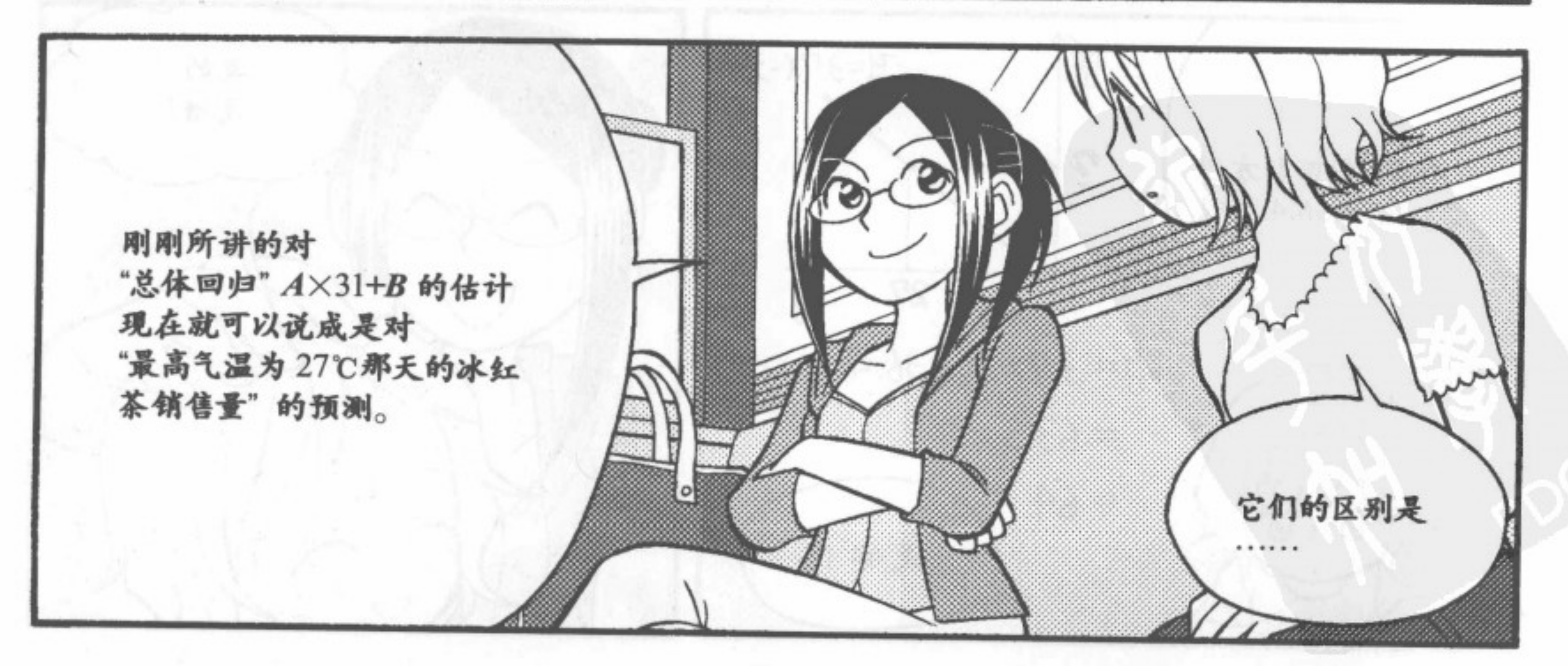
y=3.7x-36.4
?
27
嗯……刚刚求出的回归方程是 y=3.7x-36.4 所以……
y=3.7×27-36.4
=63.5
≈64
是64！

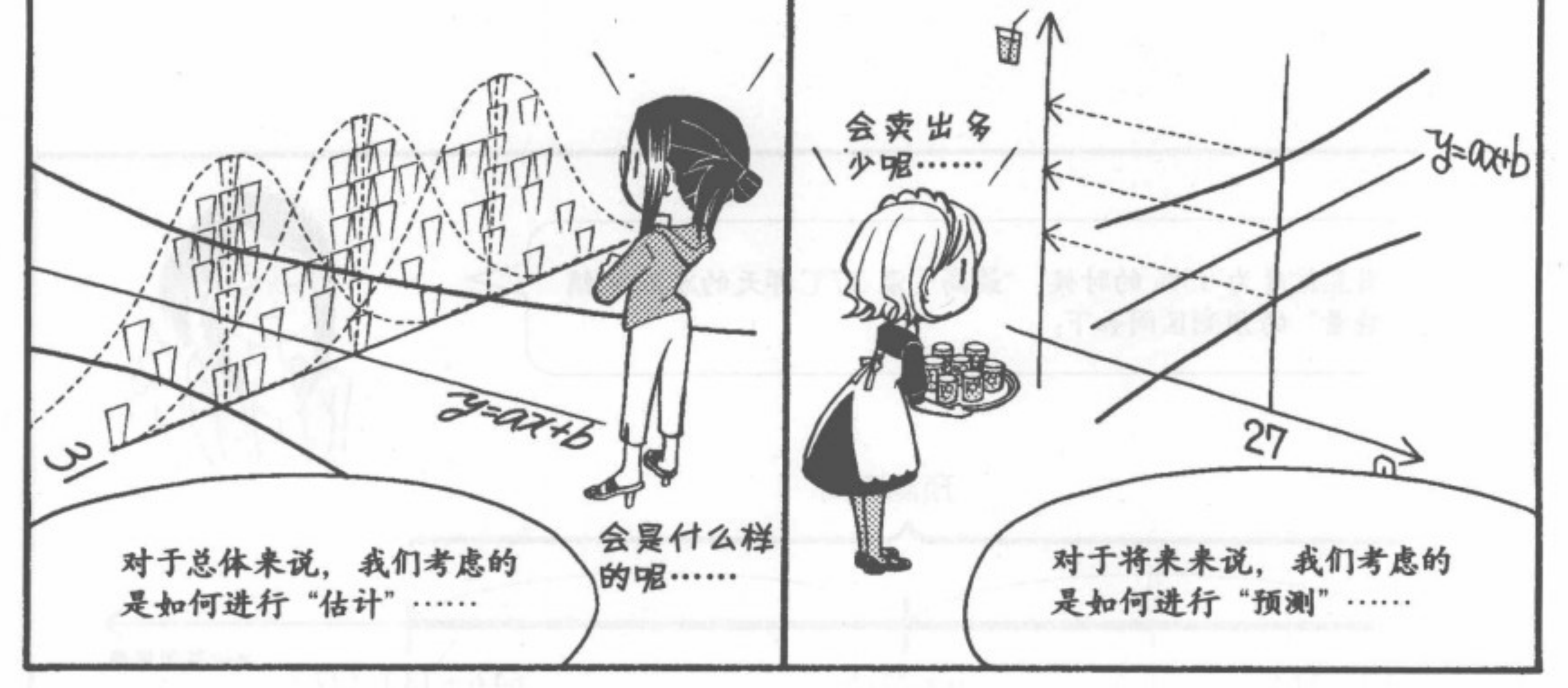


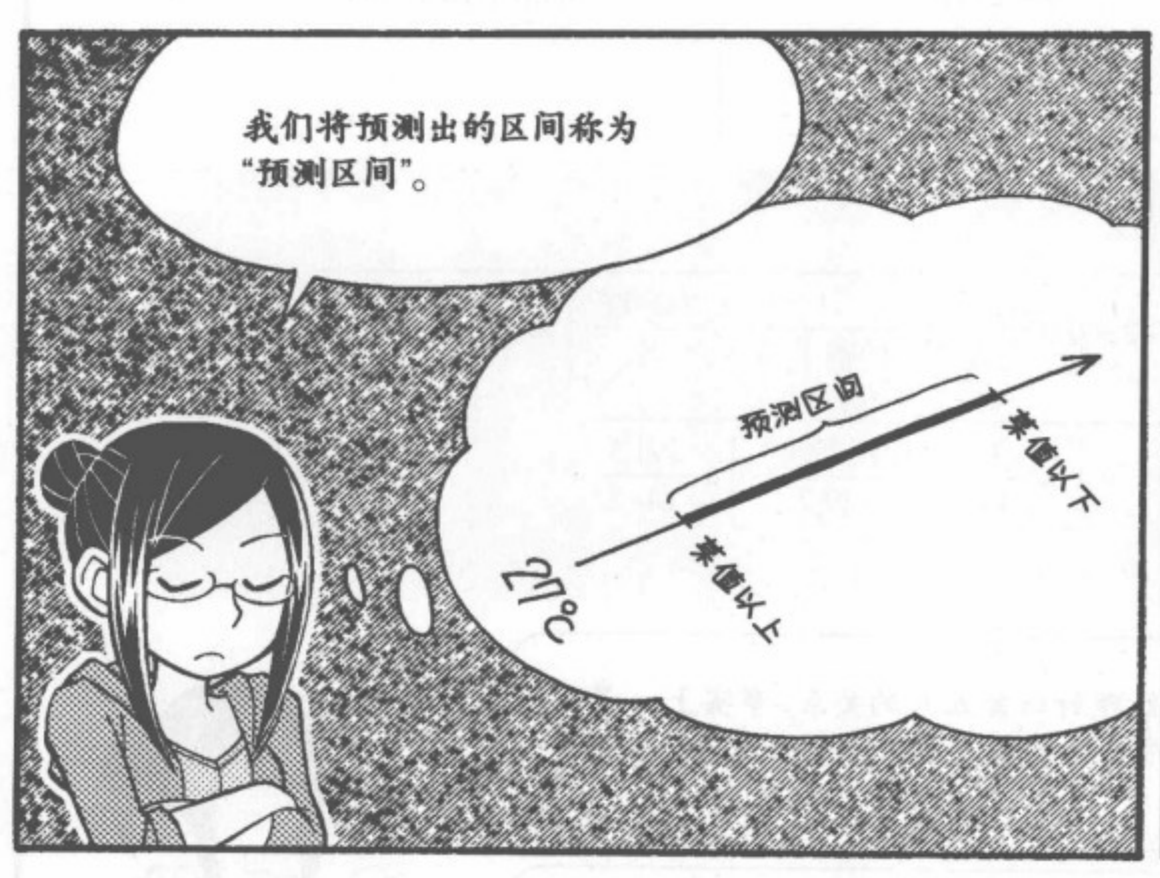
是的没错！
えへへ


但是，实际的销售量真的会刚好 64 杯吗……
我觉得，不会那么简单吧……


无论如何，都不会那么精确的！
不过，既然判定系数的值是 0.8225 那总会是 64 杯左右吧……
有一个好方法哦！
统计学中，在置信度为○○％的条件下，最高气温为 27℃那天的冰红茶销售量，会处于“某个值以上、某个值以下”还记得吗？
某个值以下
回归方程
某个值以上
27
是刚刚学过的术语……

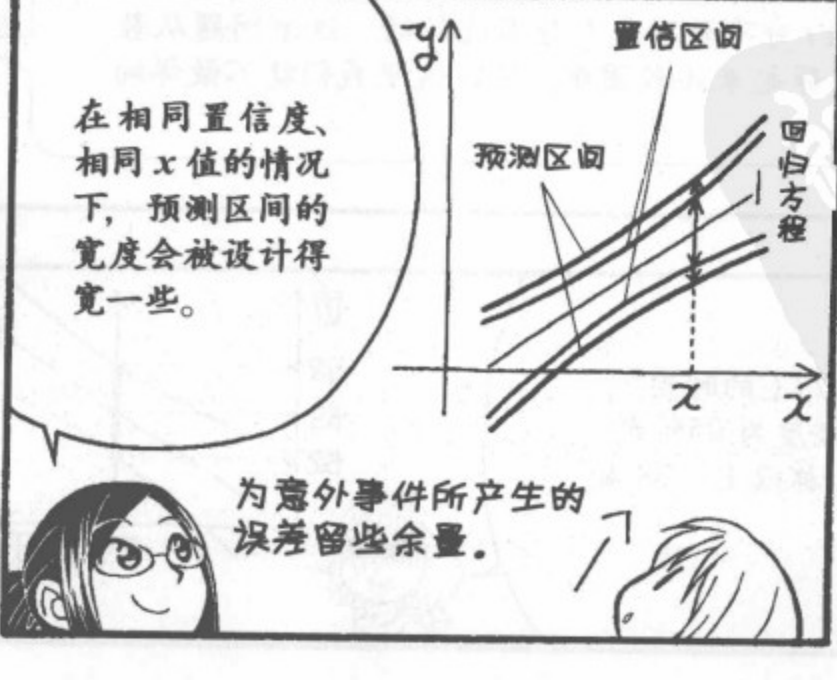

刚刚所讲的对“总体回归” A×31+B 的估计现在就可以说成是对“最高气温为 27℃那天的冰红茶销售量”的预测。
它们的区别是……


会是什么样的呢……
对于总体来说，我们考虑的是如何进行“估计”……
$y=ax+b$
会卖出多少呢……
$y=ax+b$
27
对于将来来说，我们考虑的是如何进行“预测”……


我们将预测出的区间称为“预测区间”。
预测区间
某值以下
某值以上
27℃


然后，所谓的“一定会”的可靠程度指的是“置信度”、“置信水平”或“置信系数”。
是！

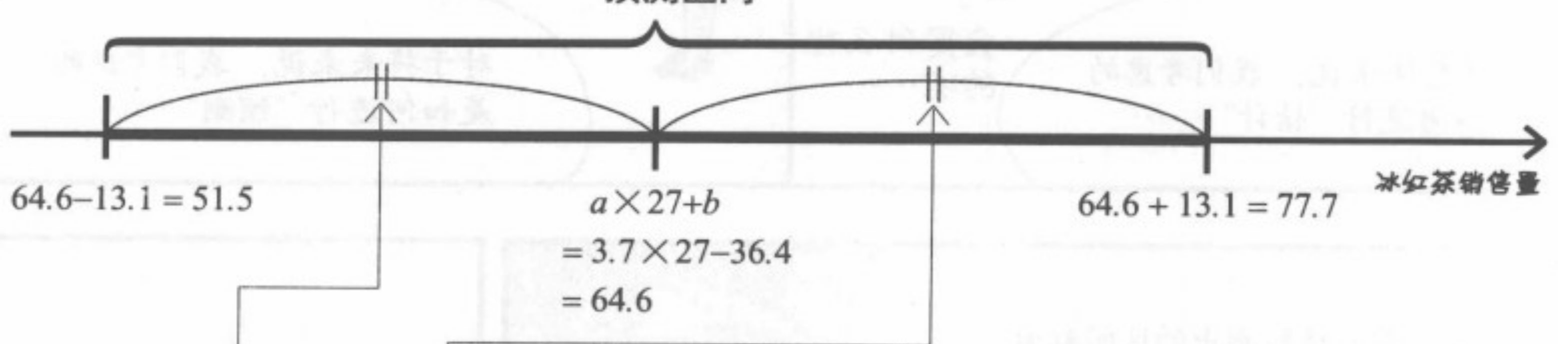

和总体回归的估计很相似啊！
计算方法也十分相似，但是有一些细微差别。


在相同置信度、相同 x 值的情况下，预测区间的宽度会被设计得宽一些。
置信区间
预测区间
回归方程
为意外事件所产生的误差留些余量。


那么求一下27℃的预测区间吧！
是！

当置信度为 95% 的时候，“最高气温 27℃那天的冰红茶销售量”的预测区间如下：





$64.6-13.1=51.5$

$a\times 27+b$

$=3.7\times 27-36.4$

$=64.6$

$64.6+13.1=77.7$

这两部分的长度都是

$$\sqrt{F(1,\ \text{样本个数}-2;\ 0.05)\times\left[1+\frac{1}{\text{样本个数}}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]\times\frac{S_e}{\text{样本个数}-2}}$$

$$=\sqrt{F(1,\ 14-2;\ 0.05)\times\left[1+\frac{1}{14}+\frac{(27-29.1)^2}{129.7}\right]\times\frac{391.1}{14-2}}$$

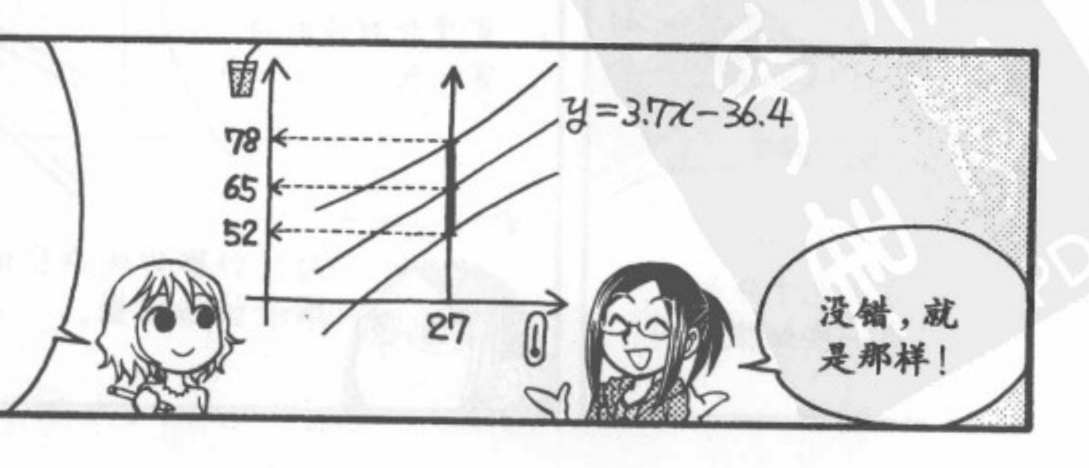
$$=13.1$$

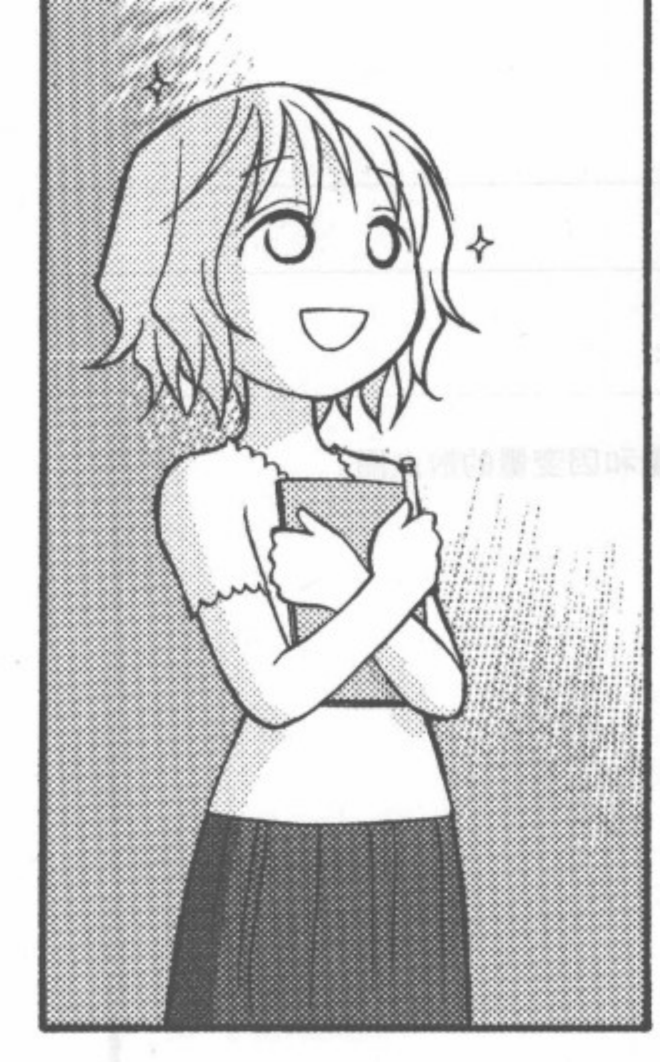
刚刚的计算中，没有处理好四舍五入的关系。事实上，最高气温为 27℃那天的冰红茶销售量，不是 64 而应当是 64.6 ≈ 65 。

通常，在区间的预测和之前所讲的总体回归的估计过程中，是根据 t 分布而不是 F 分布进行的。这个问题从数学的角度解释起来比较困难，所以这里我们就不做详细介绍了。



总之，最高气温为 27℃的时候，冰红茶销售量在置信度为 95% 的条件下，会处于 52 杯以上、78 杯以下，对吧？






快坐下！
啊呀呀！对不起，


我要讲的就这么多了。
回归分析，感觉怎么样？


嗯……
计算有些复杂，
不过还好。
通过诺伦的数据，真的学到了很多东西……

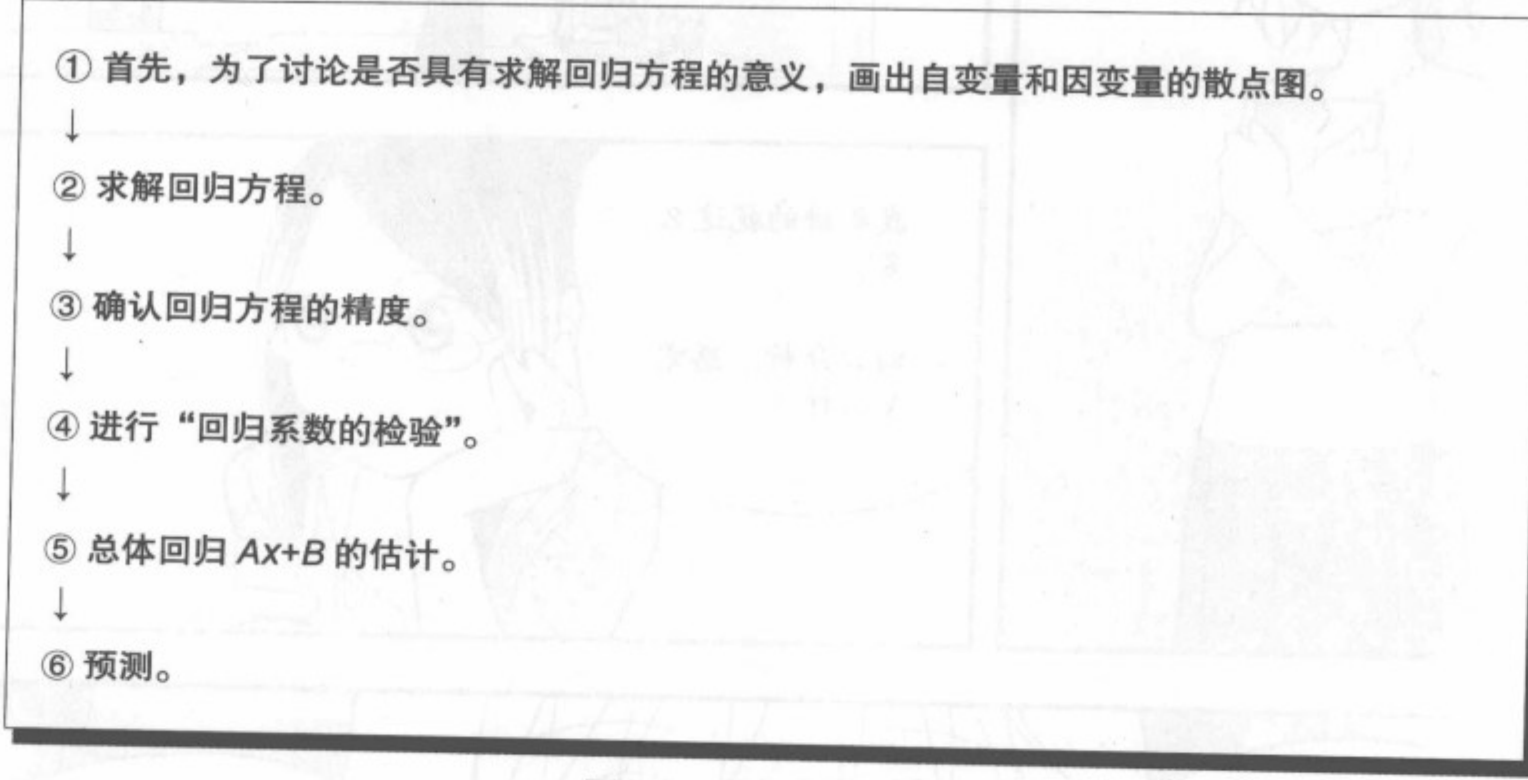

真的很感谢您啊，理纱前辈！
呵呵呵
不用客气啦！


我已经学会如何做回归分析了！


这个，有点言之过早了吧，还要及时复习啊！

✿3. 回归分析过程中的注意事项✿

下图中，我们再次给出 62 页中出现的回归分析的过程：



① 首先，为了讨论是否具有求解回归方程的意义，画出自变量和因变量的散点图。

↓

② 求解回归方程。

↓

③ 确认回归方程的精度。

↓

④ 进行“回归系数的检验”。

↓

⑤ 总体回归 $Ax+B$ 的估计。

↓

⑥ 预测。

图 2.1　回归分析的过程

此前，我们的讲解中讲到，必须完成上图中的第①步到第⑤步，但事实上并非如此。本系列的《漫画统计学》中曾提到，统计学大致可以分为两类：

- 推断统计学
- 描述统计学

那么，请再回想一下，在 25 页出现的“美羽的年龄和身高”的例子中，

- 美羽只是世界上广大人群中的一员。
- 美羽 10 岁时的身高 137.5cm 也只是其中的“一个值”。

基于以上两个事实我们可以知道，像“美羽 10 岁时的身高服从平均值为 $Ax+B$、标准差为 σ 的正态分布”等这类的问题，已经没有思考的余地了。就是说，像求解总体回归 $Ax+B$ 的置信区间，以及检验 $A \neq 0$ 是否成立这类问题，从推断统计学的观点来看，已经没有必要再进行分析了。总之，就是应当以描述统计学的观点出发进行分析。

综上所述，上图中的第①步做到第⑤步是必须要做的，请尽量掌握。但是，如果遇到像“美羽的年龄和身高”这类的情况，也就是应当从描述统计学的观点出发进行分析的情况时，分析过程只需要进行第①步做到第③步。当然如果必要的话，第⑥步也是要进行的。

✿4. 标准化残差✿

实际应用中，我们还会遇到标准化残差这一概念。所谓标准化残差就是

$$标准化残差=\frac{残差}{\sqrt{\frac{残差平方和}{样本个数-2}}}=\frac{y-\hat{y}}{\sqrt{\frac{S_e}{样本个数-2}}}$$

下表中记录的是本章例子中的标准化残差。

◆表2.1　本章例子中的标准化残差

	最高气温（℃） x	冰红茶的销售量（杯） y	冰红茶的销售量（杯） $\hat{y}=3.7x-36.4$	残差 $y-\hat{y}$	标准化残差 $\frac{y-\hat{y}}{\sqrt{\frac{391.1}{14-2}}}$
22日(一)	29	77	72.0	5.0	0.9
23日(二)	28	62	68.3	−6.3	−1.1
24日(三)	34	93	90.7	2.3	0.4
25日(四)	31	84	79.5	4.5	0.8
26日(五)	25	59	57.1	1.9	0.3
27日(六)	29	64	72.0	−8.0	−1.4
28日(日)	32	80	83.3	−3.3	−0.6
29日(一)	31	75	79.5	−4.5	−0.8
30日(二)	24	58	53.3	4.7	0.8
31日(三)	33	91	87.0	4.0	0.7
1日(四)	25	51	57.1	−6.1	−1.1
2日(五)	31	73	79.5	−6.5	−1.1
3日(六)	26	65	60.8	4.2	0.7
4日(日)	30	84	75.8	8.2	1.4

$$\frac{8.2}{\sqrt{\frac{391.1}{14-2}}}=1.4$$

标准化残差的绝对值大的个体，被看成与其他的个体性质不同。当绝对值大于3的个体存在时，应将其剔除之后再进行回归分析。

✿5. 内插法和外插法✿

下面我们再次给出由本章的例子所推导出的回归方程。

◆表2.2 “最高气温”和“冰红茶销售量”

	最高气温（℃）	冰红茶销量（杯）
22日(一)	29	77
23日(二)	28	62
24日(三)	34	93
25日(四)	31	84
26日(五)	25	59
27日(六)	29	64
28日(日)	32	80
29日(一)	31	75
30日(二)	24	58
31日(三)	33	91
1日(四)	25	51
2日(五)	31	73
3日(六)	26	65
4日(日)	30	84

$$y = 3.7x - 36.4$$

（y：冰红茶销量；x：最高气温）

从上表中可以看出，自变量“最高气温”的最小值是24℃，最大值是34℃。

实际应用中，我们还会遇到内插和外插的概念。所谓的内插，以上表为例，就是将24℃以上并且34℃以下的值代入回归方程，进而来预测冰红茶的销售量。所谓外插，以上表为例，将24℃以下或者34℃以上的值代入回归方程，进而来预测冰红茶的销售量。

在使用外插法时，一定要注意以下情况。例如，在预测“最高气温为18℃时冰红茶销售量”的时候，我们将18代入回归方程中的x，自然就可以求出结果。预测区间也可以根据第92页的计算方法求出结果。但是，这些值和区间是否可信，并不能够从数学上得到证明。

在实际操作中，用到外插法的情况也不少。对笔者本人来说，只要不是学术研究，觉得"也许无需这么较真"，那么使用外插法也是可以的。话虽如此，但是和自变量的最小值或最大值相差太远的值，使用外插法还是被认为不太可靠。

✿ 6. 序列相关 ✿

在本章的例子中，是将"最高气温"作为自变量的。那么请您想一想，某天的最高气温是 30℃，而第二天突然下降了 20℃，这种情况似乎不大可能。通常，要花上几天的时间，气温才能渐渐地降下去或者升上来，而相应的因变量"冰红茶的销售量"也只能渐渐地随之变化。

有些数据会随时间的经过而或多或少地受到影响，对这类数据的分析，我们称为"序列相关"。遇到这样的问题，最好先确认一下相邻残差之间的关联程度。有时序列相关也称为自相关。

通常，我们使用 Durbin-Watson 统计量作为衡量序列相关程度的指标。可以经过以下计算求得。

$$\text{Durbin-Watson 统计量} = \frac{\text{相邻残差的差的平方之和}}{\text{各残差的平方之和}}$$

如果这个值在 2 左右，那就说明不存在序列相关，也就是说没问题。

在本章的例子中，可以求得其 Durbin-Watson 统计量为如下：

$$\frac{(-6.3-5.0)^2+[2.3-(-6.3)]^2+\cdots+(8.2-4.2)^2}{5.0^2+(-6.3)^2+\cdots+8.2^2}=1.7$$

这个值接近 2，所以可以说不存在序列相关。

✿ 7. 直线以外的回归方程 ✿

在第 60 页中，

> 所谓回归分析，就是求出被称为回归方程的
>
> $$y=ax+b$$
>
> 的一种分析方法。

需要说明一点。实际上，所求的回归方程并不一定是 $y=ax+b$ 这样的“直线”，例如，还有如下形式：

- $y=\frac{a}{x}+b$
- $y=a\sqrt{x}+b$
- $y=ax^2+bx+c$
- $y=a\log x+b$

以上这些形式都是可以的。实际上，第 26 页出现的美羽“年龄”和“身高”的回归方程，并不是 $y=ax+b$ 的形式，而是 $y=\frac{a}{x}+b$ 的形式。

应当使用哪种函数类型来求解回归方程。这要根据分析者自己的判断进行选择。基本上，按照以下的顺序进行判断较为合理。

① 画出自变量和因变量的散点图。

② 使用和散点图形式相近的函数类型求解回归方程，例如在第 26 页的例子中，$y=\frac{a}{x}+b$ 和 $y=a\sqrt{x}+b$ 都符合，那就两个都用，将能求出来的都求出来。

③ 在第②步所求解的回归方程当中，判定系数的值较大的那个就是我们所要求的回归方程。

第 26 页中出现的美羽的“年龄”和“身高”的回归方程，$y=-\frac{326.6}{x}+173.3$ 是如何求出来的呢？下面给出计算过程。

■ 美羽“年龄”和“身高”的回归方程的求解方法

$y=\frac{a}{x}+b$，令 $\frac{1}{x}=X$

$$y=\frac{a}{x}+b=aX+b$$

这样就可以将其改写成“直线”的形式。

正如第 70 页说明的那样，回归方程 $y=aX+b$ 中，a 和 b 的值可以通过以下计算求得。

$$\begin{cases} a=\frac{S_{Xy}}{S_{XX}} \\ b=\bar{y}-\bar{X}a \end{cases}$$

所以，根据下一页中的表 2.3 可知

$$\begin{cases} a=\frac{S_{Xy}}{S_{XX}}=\frac{-15.9563}{0.0489}=-326.6^{※} \\ b=\bar{y}-\bar{X}a=138.2625-0.1072\times(-326.6)=173.3 \end{cases}$$

因此，回归方程为：$y=-326.6x+173.3$ 于是 $y=-\frac{326.6}{x}+173.3$

（y：身高；x：$\frac{1}{年龄}$）　（y：身高；x：年龄）

※这里所写的计算值不应是 −326.6，而应是所求得的 − 326.3。这是由于四舍五入而产生的误差。

◆表 2.3 a和 b的计算过程

	年龄 x	$\frac{1}{年龄}$ $\frac{1}{x}=X$	身高 y	$X-\bar{X}$	$y-\bar{y}$	$(X-\bar{X})^2$	$(y-\bar{y})^2$	$(X-\bar{X})(y-\bar{y})^2$
	4	0.2500	100.1	0.1428	−38.1625	0.0204	1456.3764	−5.4515
	5	0.2000	107.2	0.0928	−31.0625	0.0086	964.8789	−2.8841
	6	0.1667	114.1	0.0595	−24.1625	0.0035	583.8264	−1.4381
	7	0.1429	121.7	0.0357	−16.5625	0.0013	274.3164	−0.5914
	8	0.1250	126.8	0.0178	−11.4625	0.0003	131.3889	−0.2046
	9	0.1111	130.9	0.0040	−7.3625	0.0000	54.2064	−0.0292
	10	0.1000	137.5	−0.0072	−0.7625	0.0001	0.5814	0.0055
	11	0.0909	143.2	−0.0162	4.9375	0.0003	24.3789	−0.0802
	12	0.0833	149.4	−0.0238	11.1375	0.0006	124.0439	−0.2653
	13	0.0769	151.6	−0.0302	13.3375	0.0009	177.8889	−0.4032
	14	0.0714	154.0	−0.0357	15.7375	0.0013	247.6689	−0.5622
	15	0.0667	154.6	−0.0405	16.3375	0.0016	266.9139	−0.6614
	16	0.0625	155.0	−0.0447	16.7375	0.0020	280.1439	−0.7473
	17	0.0588	155.1	−0.0483	16.8375	0.0023	283.5014	−0.8137
	18	0.0556	155.3	−0.05161	17.0375	0.0027	290.2764	−0.8790
	19	0.0526	155.7	−0.0545	17.4375	0.0030	304.0664	−0.9507
总计	184	1.7144	221 2.2	0.0000	0.0000	0.0489	5464.4575	−15.9563
平均	11.5	0.1072	138.3					
		↓ $\bar{X}$	↓ $\bar{y}$			↓ S_{XX}	↓ S_{yy}	↓ S_{Xy}

◆第 3 章◆

重回归分析



本章中关于“重回归分析中的预测区间”的相关内容已经传送到网站 http://www.okbook.com.cn，名为“重回归分析中的预测区间”的文件夹中，请您下载参考。



1. 重回归分析的定义
谢谢你的资料！
好啦！不用客气！说起来，像你这样用心指导学妹还真是伟大啊！
没什么……
有很多内情的……
理纱前辈……
はあっ
哦！快来快来！
这边！
はあっ
对，对不起！
拖了很久才下课……
没关系！反正没等多久。


ニコニコ
那……
这位是？
啊，介绍一下。

我的同学风见君。
请多关照！
初，
初次见面……

今天我们要讲“重回归分析”，所用的分析材料是风见君带给我们的。
太感谢您了！
哎呀！
我也是举手之劳罢了。

美羽，是不是喜欢牛角面包来着？
Croissant
是啊，非常喜欢，怎么了？

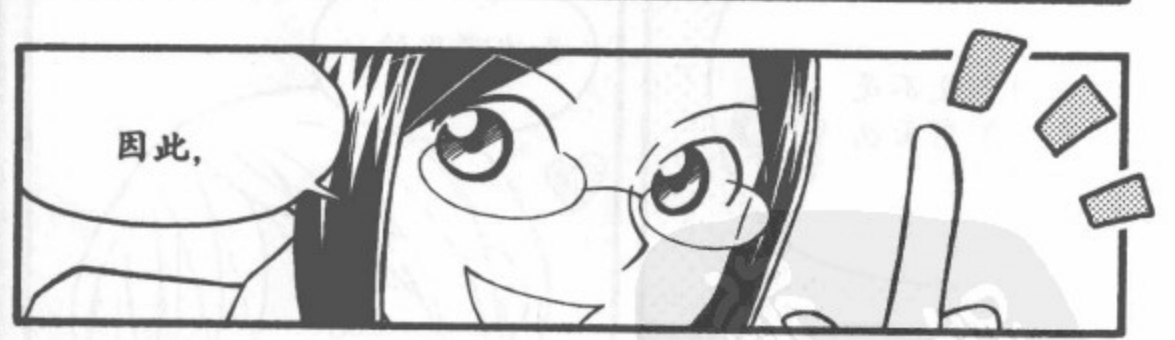
喜欢哪里的？
说起来还是风见面包店的最好吃！

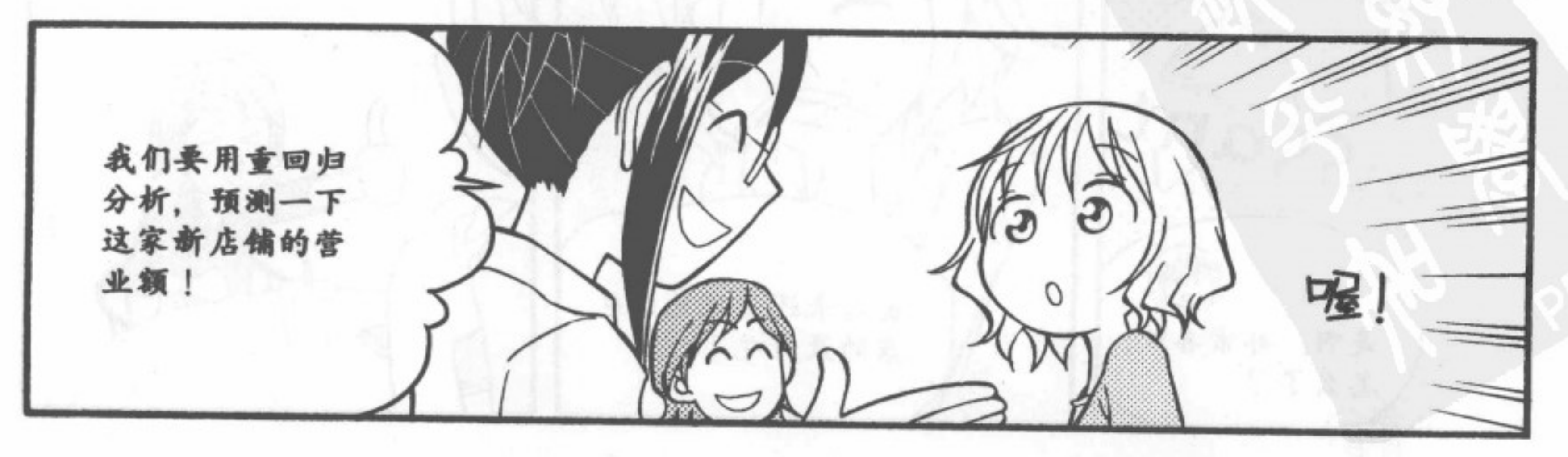
风见……


啊！
风见前辈就是……
没错！
少东家。
宫野，这个称呼
太奇怪了吧！

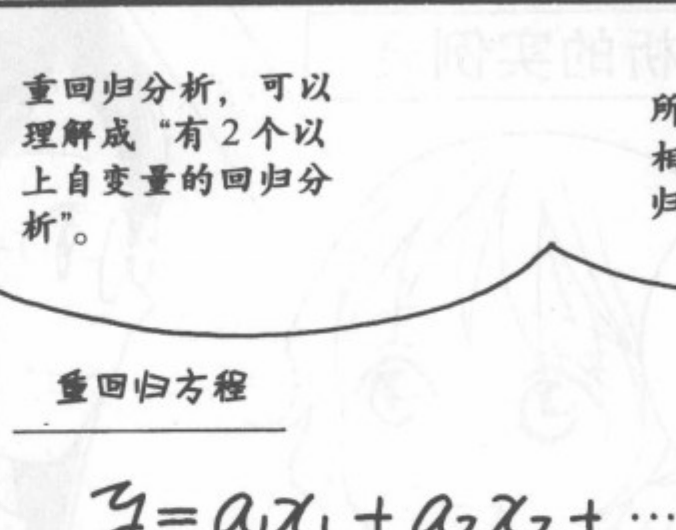

没什么了不起的。
所以……
唉，可是，我看到
过很多家分店！

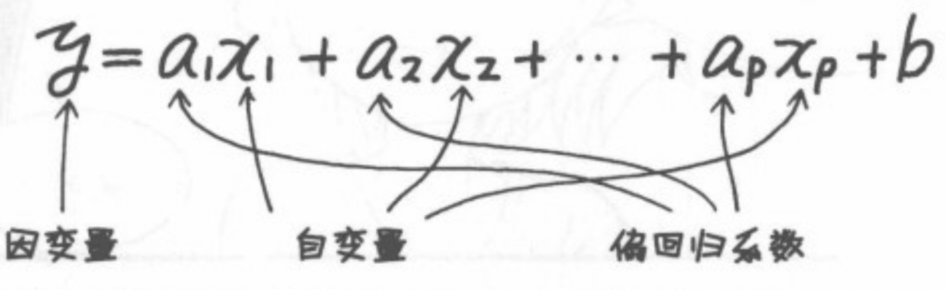

到目前为止，
有 10 个店铺。
梦之丘总店
寺井站大厦店
曾根店
桔梗町店
邮政局前店
水道町站前店
六条站大厦店
桥本大街店
若叶川店
美里店
NEW
伊势桥店
噢，现在要开第 11
家分店……

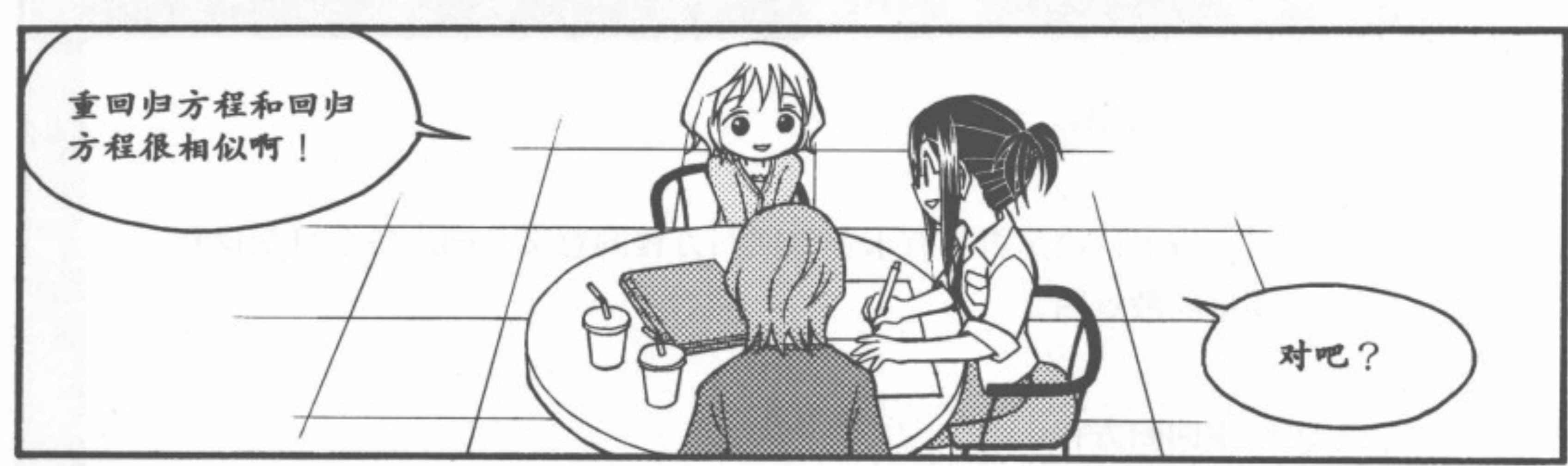

因此，

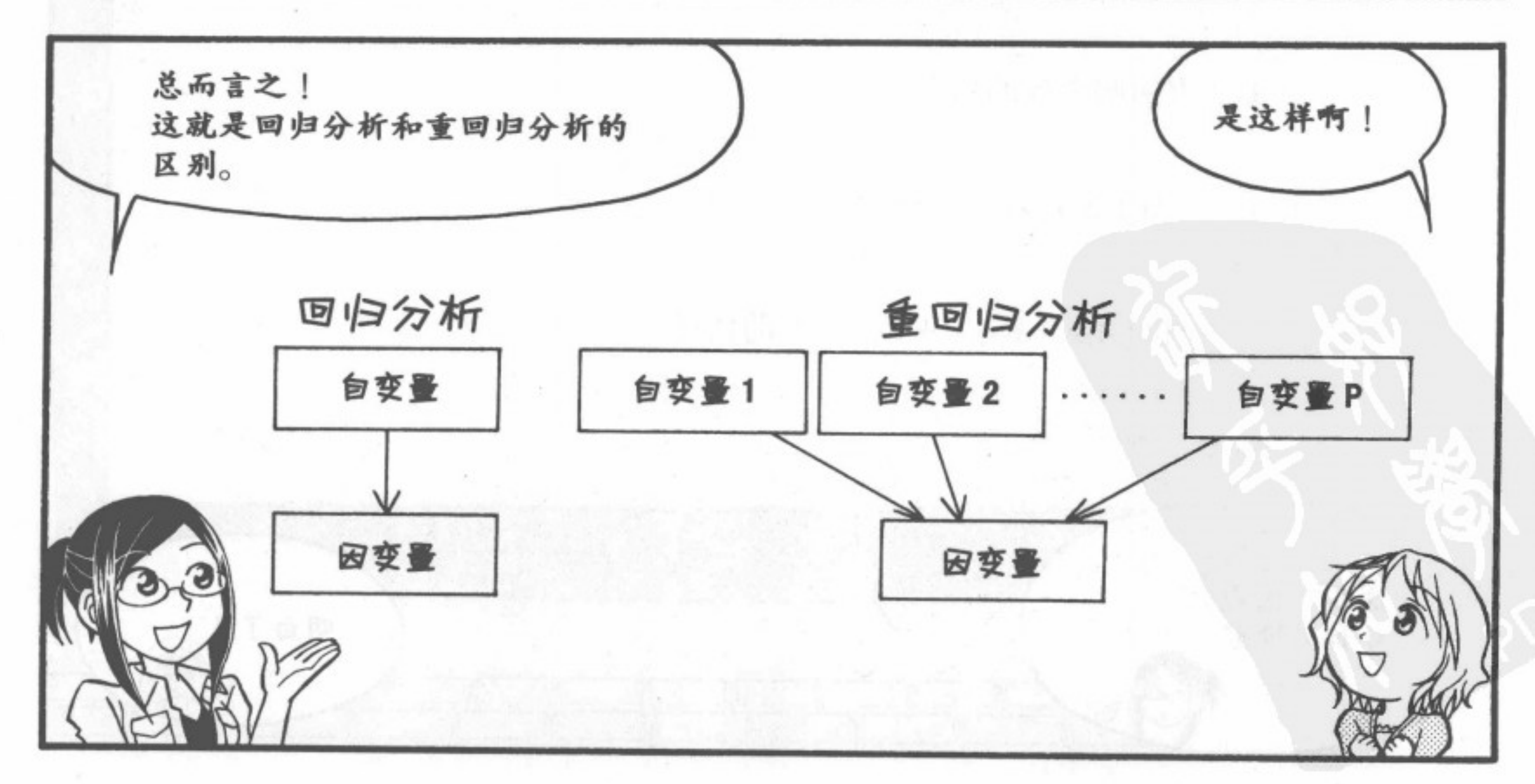

我们要用重回归
分析，预测一下
这家新店铺的营
业额！
喔！


重回归分析就是通过多个因素进行预测的一种手段。
没错！

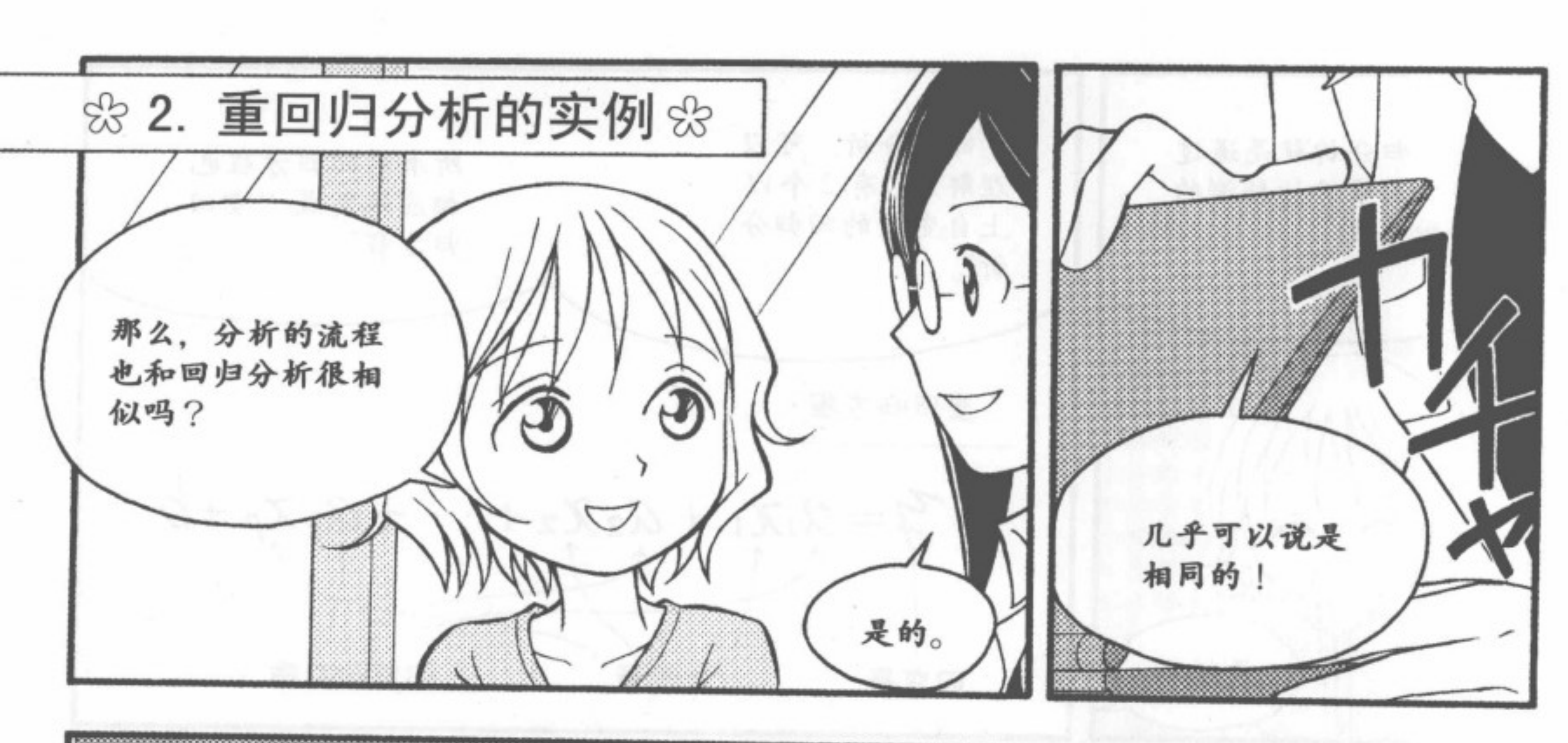

重回归分析，可以理解成"有2个以上自变量的回归分析"。
所求的回归方程也相应地变成"重回归方程"！


重回归方程
$y = a_1x_1 + a_2x_2 + \cdots + a_px_p + b$
因变量
自变量
偏回归系数


重回归方程和回归方程很相似啊！
对吧？


总而言之！
这就是回归分析和重回归分析的区别。
是这样啊！
回归分析
自变量
因变量
重回归分析
自变量1
自变量2
……
自变量P
因变量

✿ 2. 重回归分析的实例 ✿



重回归分析的流程

① 首先，为了讨论是否具有求解重回归方程的意义，画出各个自变量和因变量的散点图。

↓

② 求解重回归方程。

↓

③ 确认重回归方程的精度。

↓

④ 进行“重回归系数的检验”。

↓

⑤ 总体回归 $A_1x_1+A_2x_2+\cdots+A_px_p+B$ 的估计。

↓

⑥ 预测。

这就是重回归分析的过程。

明白了！

① 首先，为了讨论是否具有求解重回归方程的意义，画出各个自变量和因变量的散点图

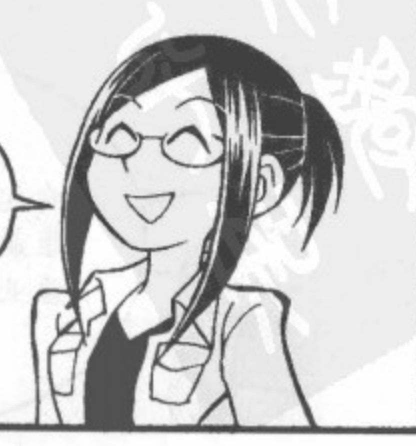


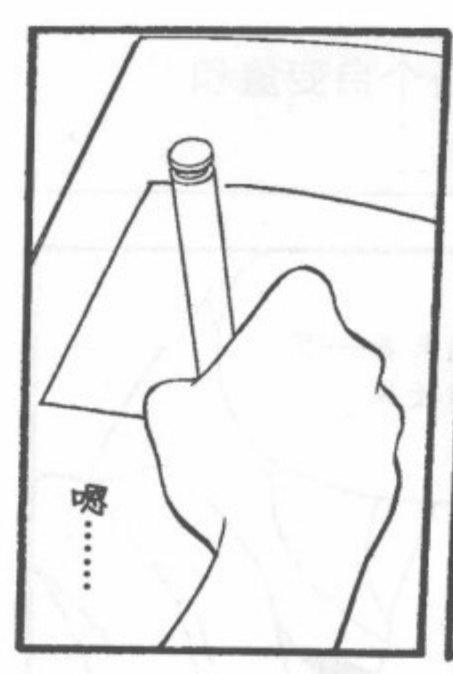


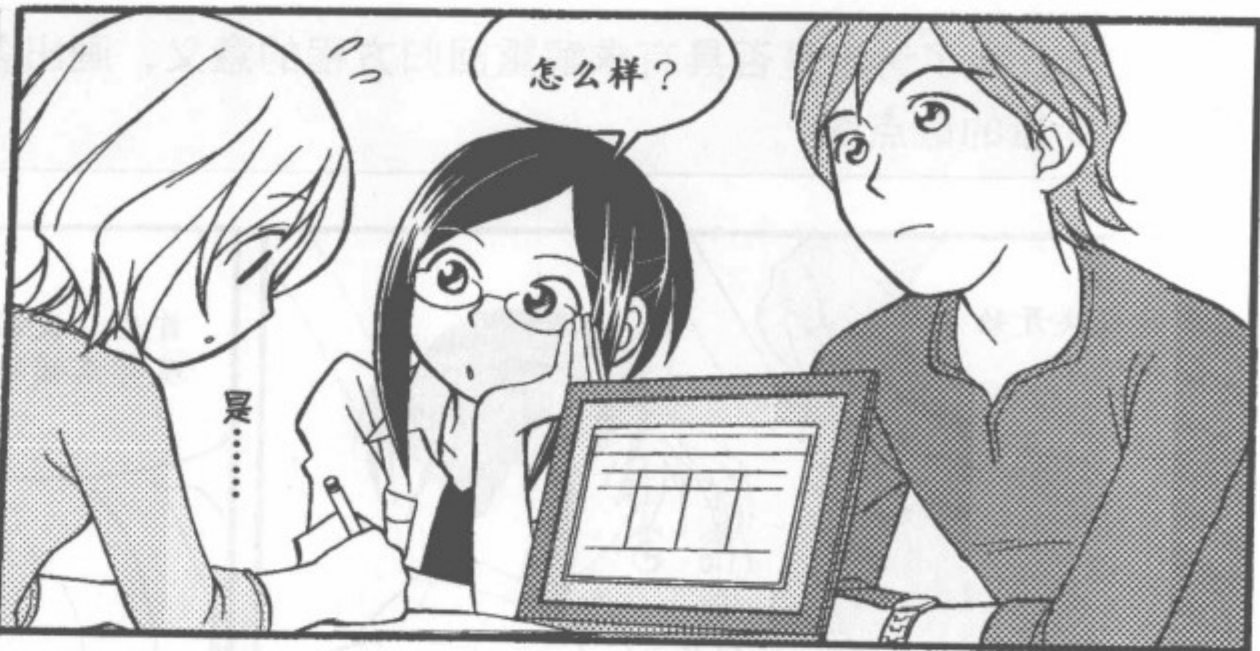
	店铺的面积（坪※）	距离最近的车站（m）	月营业额（万日元）
梦之丘总店	10	80	469
寺井站大厦店	8	0	366
曾根店	8	200	371
桥本大街店	5	200	208
桔梗町店	7	300	246
邮政局前店	8	230	297
水道町站前店	7	40	363
六条站大厦店	9	0	436
若叶川店	6	330	198
美里店	9	180	364

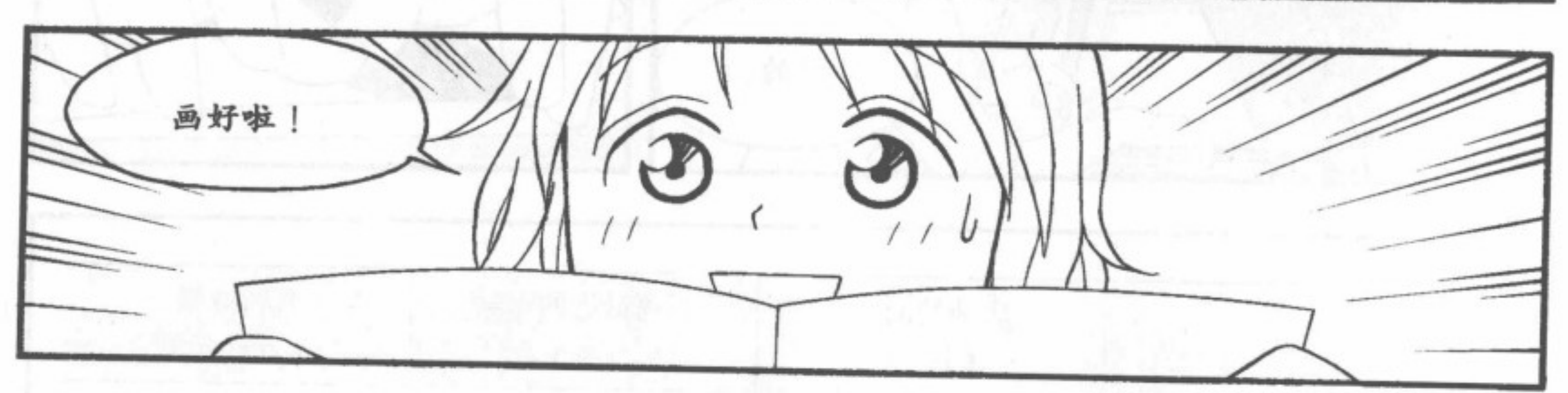
※坪：日本面积单位名，1坪约为3.305785m²。











“店铺面积”和“月营业额”
单相关系数＝0.8924

月营业额（0–500）；店铺面积（2–12）

梦之丘总店
六条站大厦店
曾根店
美里店
水道町站前店
寺井站大厦店
邮政局前店
桥本大街店
桔梗町店
若叶川店

“距最近车站的距离”和“月营业额”
单相关系数＝−0.7751

月营业额（0–500）；距最近车站的距离（50–350）

梦之丘总店
六条站大厦店
寺井站大厦店
曾根店
水道町站前店
美里店
邮政局前店
桔梗町店
桥本大街店
若叶川店

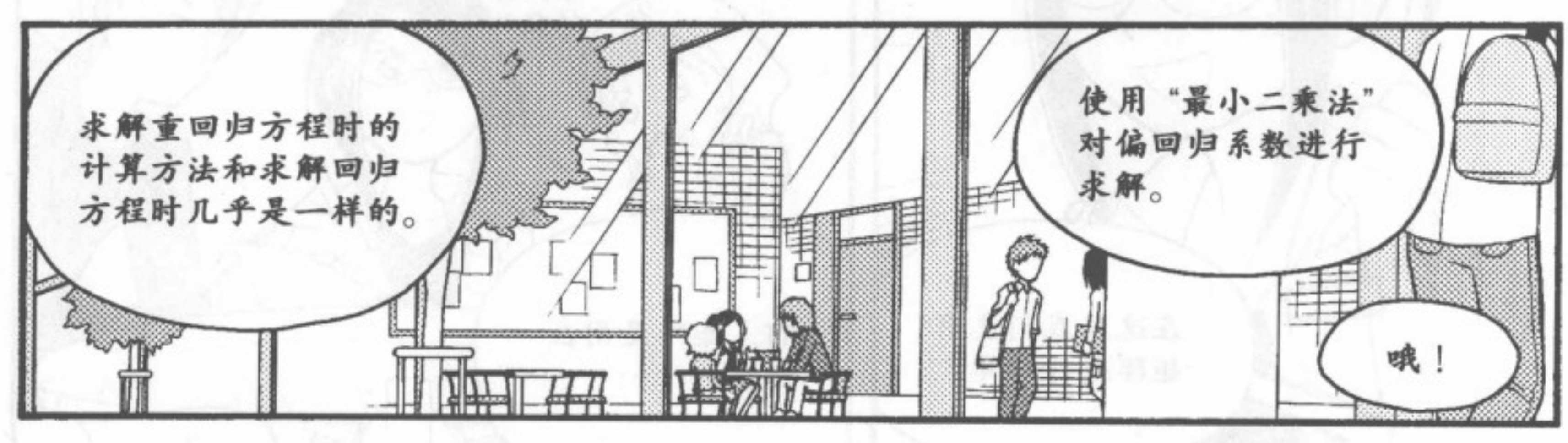
嗯，

看起来，店铺面积越大，距最近的车站越近，月营业额就越高！

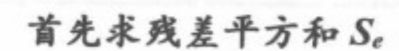
具有求解重回归方程的意义啊！

是啊！

② 求解重回归方程







$$S_e=\{469-(a_1\times10+a_2\times80+b)\}^2+\{366-(a_1\times8+a_2\times0+b)\}^2+\cdots+\{364-(a_1\times9+a_2\times180+b)\}^2$$



其次，关于 a_1、a_2 和 b 求微分，令微分值为 0，再求出令 S_e 的值最小时的 a_1、a_2 和 b 的值……

$$\frac{dS_e}{da_1}=2(-10)\{469-(a_1\times10+a_2\times80+b)\}+2(-8)\{366-(a_1\times8+a_2\times0+b)\}+\cdots+2(-9)\{364-(a_1\times9+a_2\times180+b)\}=0$$

$$\frac{dS_e}{da_2}=2(-80)\{469-(a_1\times10+a_2\times80+b)\}+2(-0)\{366-(a_1\times8+a_2\times0+b)\}+\cdots+2(-180)\{364-(a_1\times9+a_2\times180+b)\}=0$$

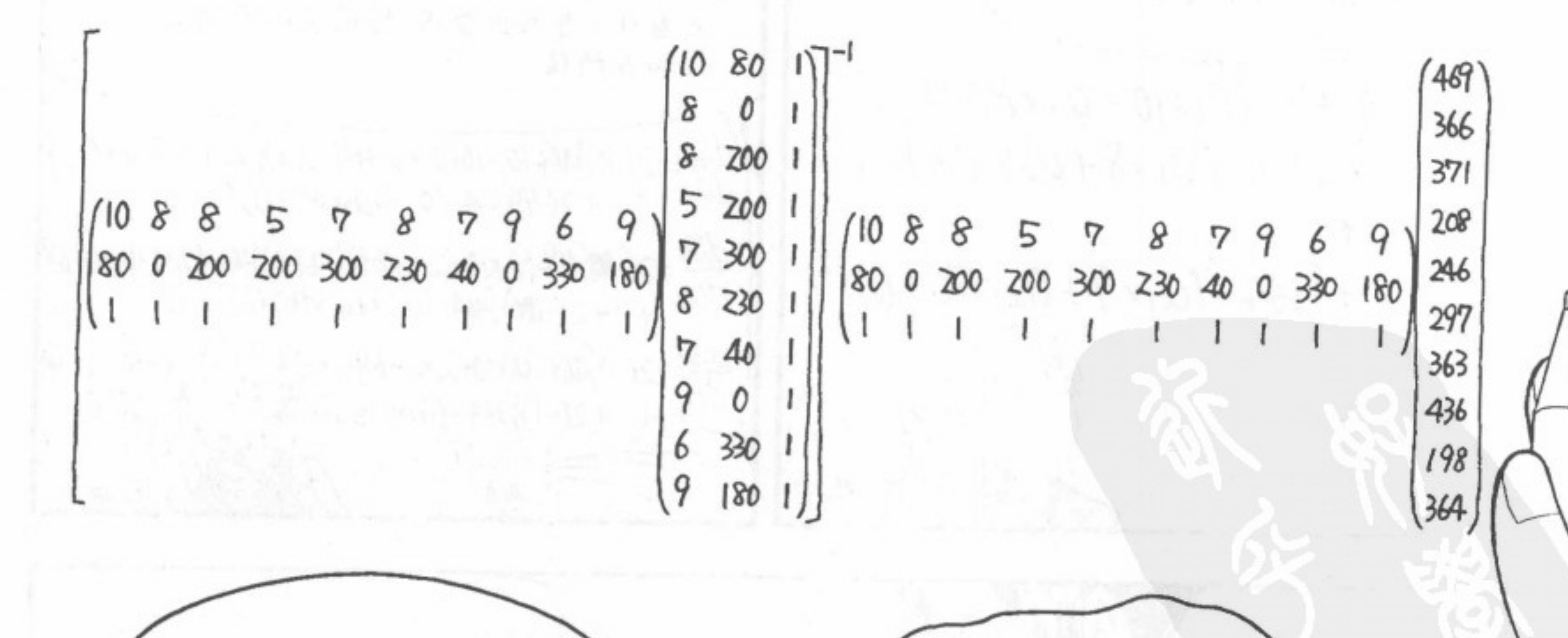
$$\frac{dS_e}{db}=2(-1)\{469-(a_1\times10+a_2\times80+b)\}+2(-1)\ldots+a_2\times0+b)\}+\cdots+2(-1)\{364-(a_1\times9\ldots$$



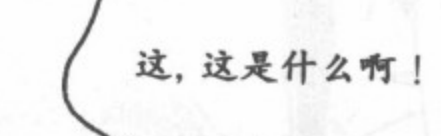








用这样的计算
就能求出来啦！



$$
\left[
\begin{pmatrix}
10 & 8 & 8 & 5 & 7 & 8 & 7 & 9 & 6 & 9 \\
80 & 0 & 200 & 200 & 300 & 230 & 40 & 0 & 330 & 180 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
10 & 80 & 1 \\
8 & 0 & 1 \\
8 & 200 & 1 \\
5 & 200 & 1 \\
7 & 300 & 1 \\
8 & 230 & 1 \\
7 & 40 & 1 \\
9 & 0 & 1 \\
6 & 330 & 1 \\
9 & 180 & 1
\end{pmatrix}
\right]^{-1}
\begin{pmatrix}
10 & 8 & 8 & 5 & 7 & 8 & 7 & 9 & 6 & 9 \\
80 & 0 & 200 & 200 & 300 & 230 & 40 & 0 & 330 & 180 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}
\begin{pmatrix}
469 \\ 366 \\ 371 \\ 208 \\ 246 \\ 297 \\ 363 \\ 436 \\ 198 \\ 364
\end{pmatrix}
$$

距最近车站的距离

店铺面积

为了便于计算，在求解重回归方程的时候，必须将这里都写成1。

月营业额

各个数值代表的意义是这样的。







自变量	偏回归系数	
店铺面积	a_2	41.5
距最近车站的距离	a_2	−0.3
常数项	b	65.3











③ 确认重回归方程的精度



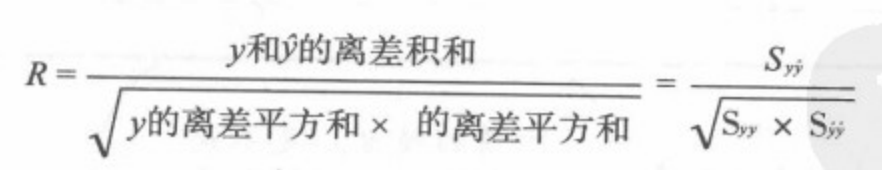
实测值 y 和预测值 $\hat{y}$ 的单相关系数就是重相关系数 R，将其平方之后就得到判定系数 R^2。



	实测值 y	预测值 $\hat{y}=41.5x_1-0.3x_2+65.3$	$y-\bar{y}$	$\hat{y}-\bar{\hat{y}}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{\hat{y}})^2$	$(y-\bar{y})(\hat{y}-\bar{\hat{y}})$	$(y-\hat{y})^2$
梦之丘总店	469	453.2	137.2	121.4	18823.8	14735.1	16654.4	250.0
寺井站大厦店	366	397.4	34.2	65.6	1169.6	4307.5	2244.6	988.0
曾根店	371	329.3	39.2	−2.5	1536.6	6.5	−99.8	1742.6
桥本大街店	208	204.7	−123.8	−127.1	15326.4	16150.7	15733.2	10.8
桔梗町店	246	253.7	−85.8	−78.1	7361.6	6106.9	6705.0	58.6
邮政局前店	297	319.0	−34.8	−12.8	1211.0	163.1	444.4	485.3
水道町站前店	363	342.3	31.2	10.5	973.4	109.9	327.1	429.2
六条站大厦店	436	438.9	104.2	107.1	10857.6	11480.1	11164.5	8.7
若叶川店	198	201.9	−133.8	−129.9	17902.4	16870.5	17378.8	15.3
美里店	364	377.6	32.2	45.8	1036.8	2096.4	1474.3	184.6
总　计	3318	3318	0	0	76199.6	72026.6	72026.6	4173.0
平　均	331.8	331.8						
	↓ $\bar{y}$	↓ $\bar{\hat{y}}$			↓ S_{yy}	↓ $S_{\hat{y}\hat{y}}$	↓ $S_{y\hat{y}}$	↓ S_e

S_e 在计算重相关系数 R 时，虽然不会涉及，但是对之后的运算很重要，所以要先求出来。

重相关系数 R 是

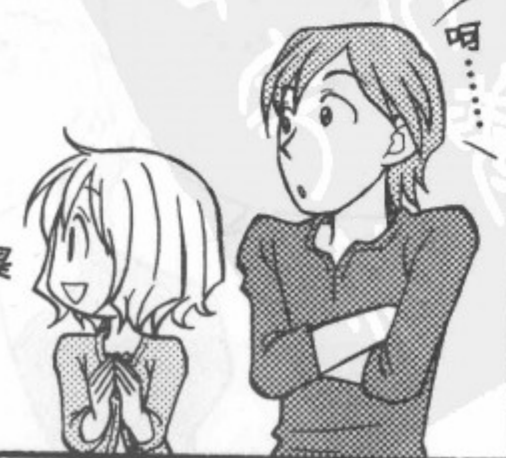


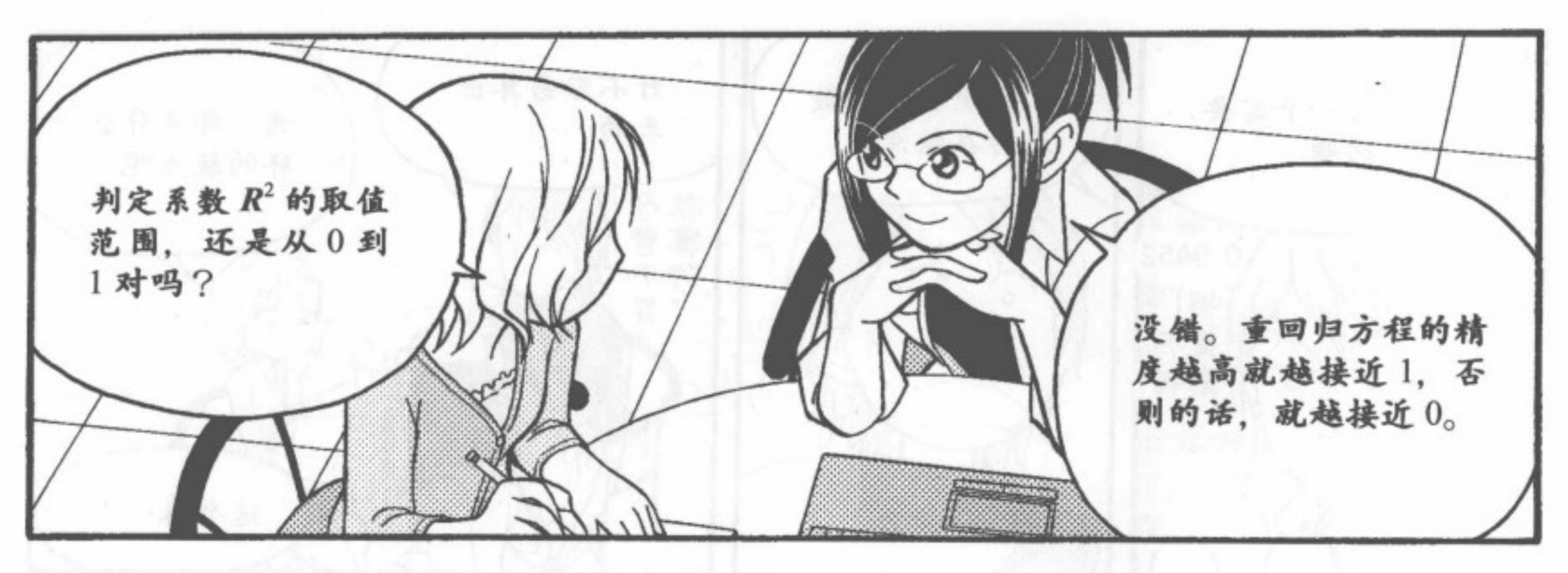
$$R=\frac{y\text{和}\hat{y}\text{的离差积和}}{\sqrt{y\text{的离差平方和}\times\ \text{的离差平方和}}}=\frac{S_{y\hat{y}}}{\sqrt{S_{yy}\times S_{\hat{y}\hat{y}}}}$$

$$=\frac{72026.6}{\sqrt{76199.6\times 72026.6}}=0.9722$$

判定系数 R^2 是

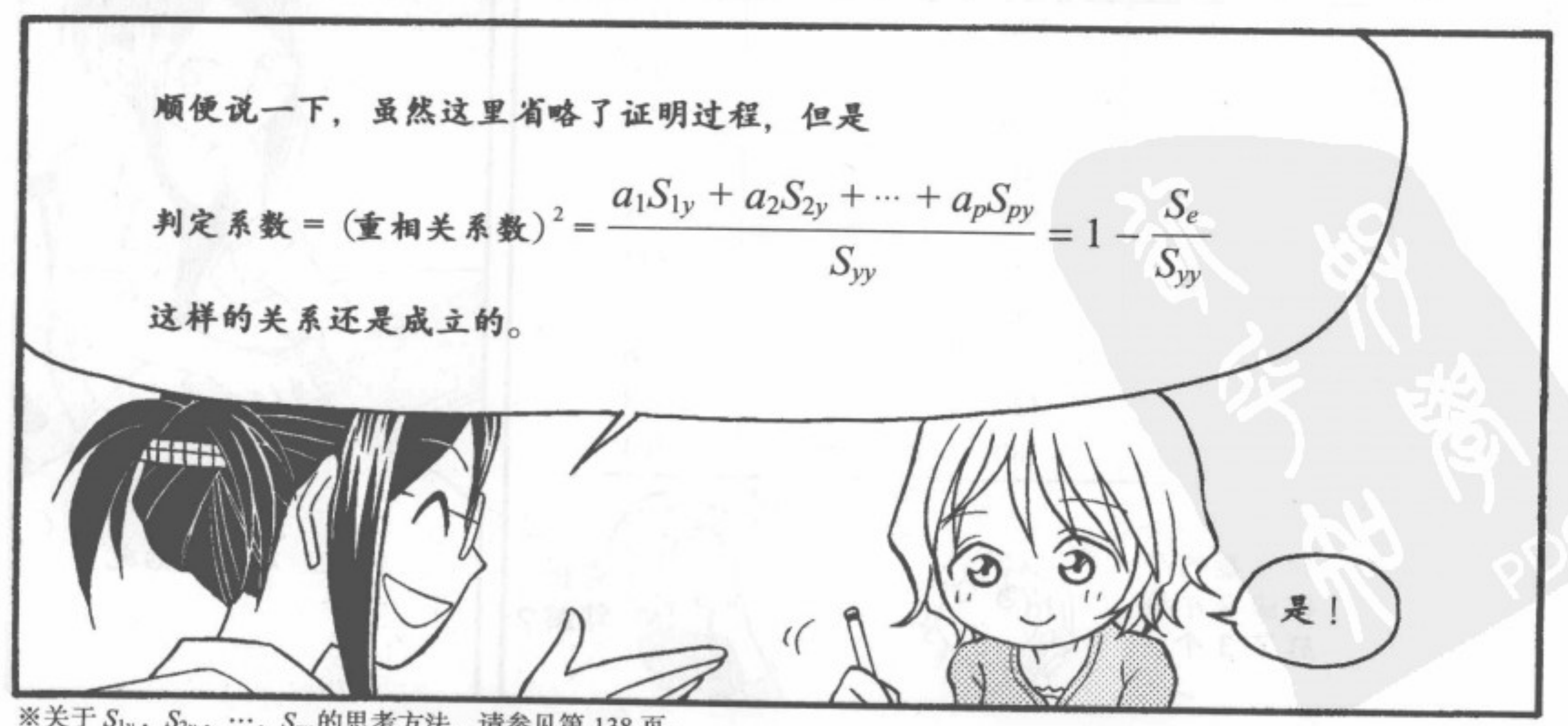
$$R^2=(0.9722)^2=0.9452$$









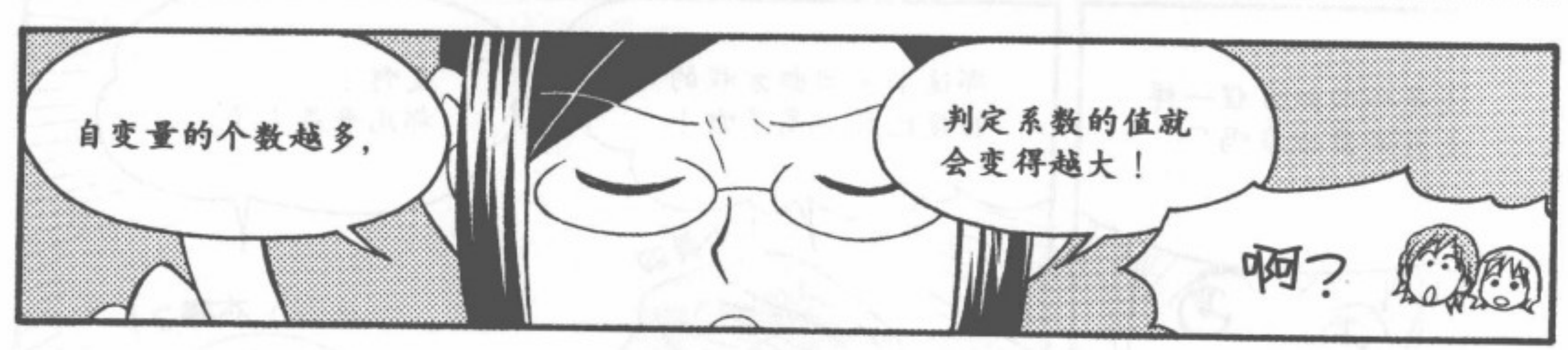


※关于 S_{1y}，S_{2y}，…，S_{py} 的思考方法，请参见第 138 页。









例如，现在我们将数据中加入“店长年龄”这一项来看看。

	店铺的面积（坪）	距最近车站的距离（m）	店长年龄（岁）	月营业额（万日元）
梦之丘总店	10	80	42	469
寺井站大厦店	8	0	29	366
曾根店	8	200	33	371
桥本大街店	5	200	41	208
桔梗町店	7	300	33	246
邮政局前店	8	230	35	297
水道町站前店	7	40	40	363
六条站大厦店	9	0	46	436
若叶川店	6	330	44	198
美里店	9	180	34	364







店铺的面积 坪
距最近的车站的距离
店长年龄

$R^2=0.9495$ > $R^2=0.9452$

店铺的面积 坪
距最近的车站的距离
店长年龄

变成了 0.9495！

真的，变大了呀！

然而，必须注意的是，

单相关系数 =0.0368
月营业额
500
400
300
200
100
0
0 10 20 30 40 50
店长年龄
梦之丘总店
六条站大厦店
曾根店
寺井站大厦店
美里店
水道町站前店
邮政局前店
桔梗町店
桥本大街店
若叶川店

"店长年龄"和"月营业额"是完全没有关系的！

我刚刚也觉得奇怪呢！

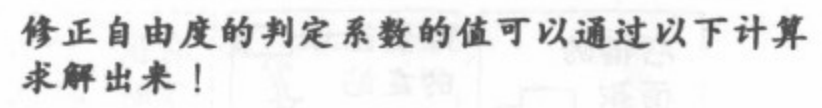
尽管如此，判定系数的值还是变大了……

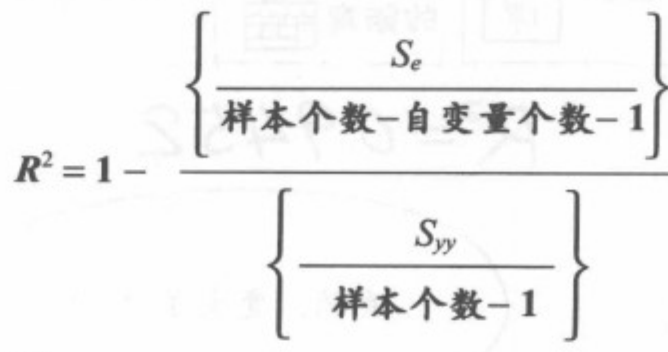
这就难办了！

但是尽管放心！

这里还有"修正自由度的判定系数"！

好恐怖的名字！



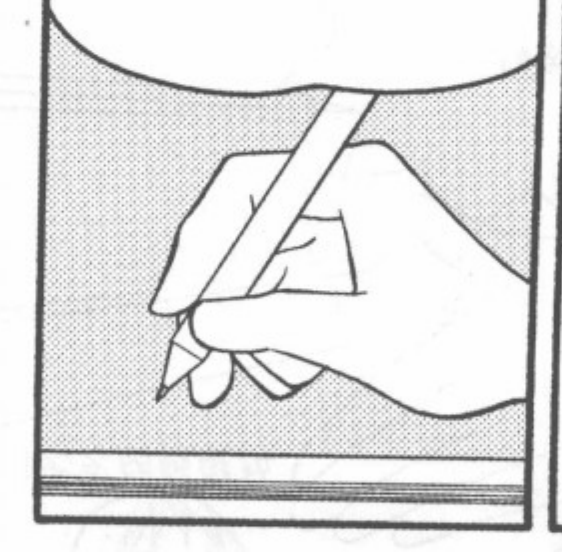


$$R^2 = 1 - \frac{\left\{\dfrac{S_e}{\text{样本个数}-\text{自变量个数}-1}\right\}}{\left\{\dfrac{S_{yy}}{\text{样本个数}-1}\right\}}$$









① 只有“店铺面积”和“距最近车站的距离”的情况

· 判定系数 R^2 是 0。9452

· 修正自由度的判定系数 R^{*2} 是

$$1 - \frac{\left(\dfrac{S_e}{\text{样本个数}-\text{自变量个数}-1}\right)}{\left(\dfrac{S_{yy}}{\text{样本个数}-1}\right)}$$

$$= 1 - \frac{\left(\dfrac{4173.0}{10-2-1}\right)}{\left(\dfrac{76199.6}{10-1}\right)} = \underline{0.9296}$$

是这样啊！



那么，接下来是“店铺面积”、“距最近车站的距离”以及“店长年龄”都考虑的情况……

不过，理纱前辈刚刚好像已经求出判定系数R^2的值了。

店铺的面积 坪

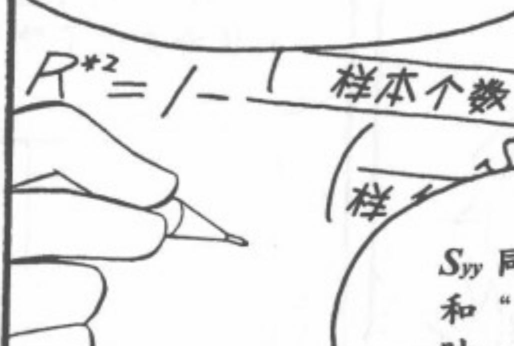
距最近车站的距离

店长的年龄

$R^2=0.9495$

对了，是0.9495！







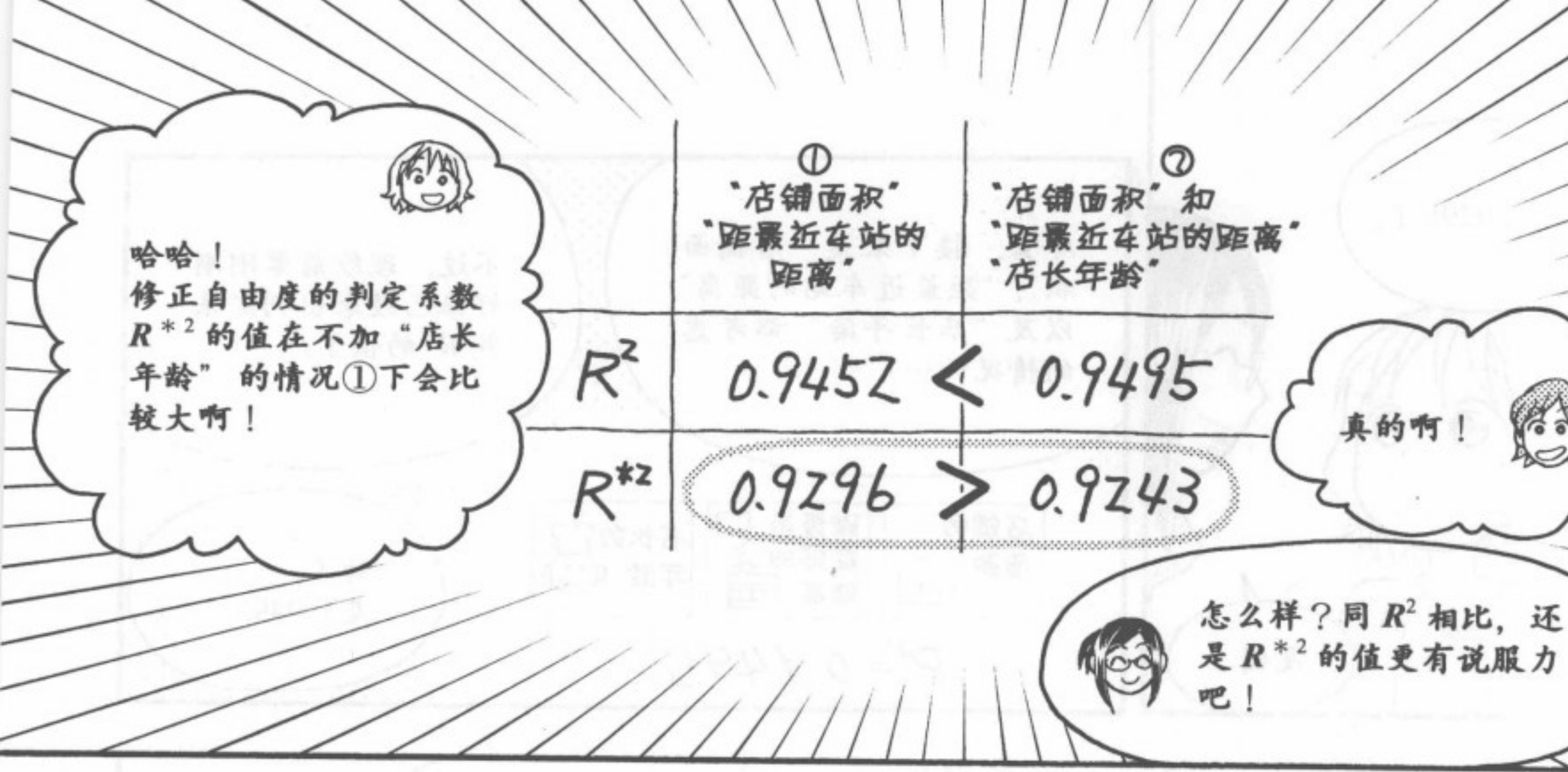
⑦“店铺面积”、“距最近车站的距离”以及“店长年龄”都考虑的情况。

- 判定系数R^2是0.9495
- 修正自由度的判定系数R^{*2}是

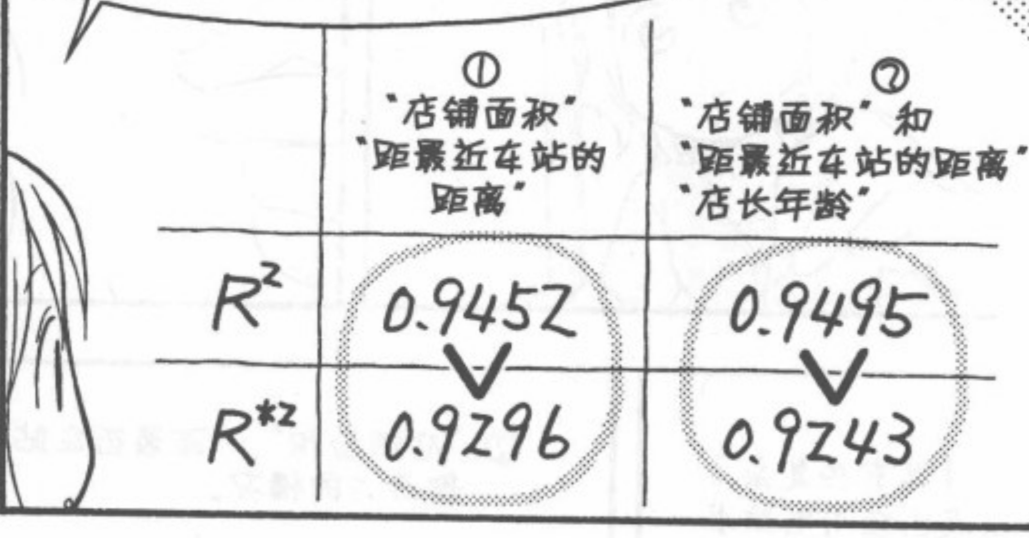
$$1-\frac{\left(\dfrac{S_e}{\text{样本个数}-\text{自变量个数}-1}\right)}{\left(\dfrac{S_{yy}}{\text{样本个数}-1}\right)}$$

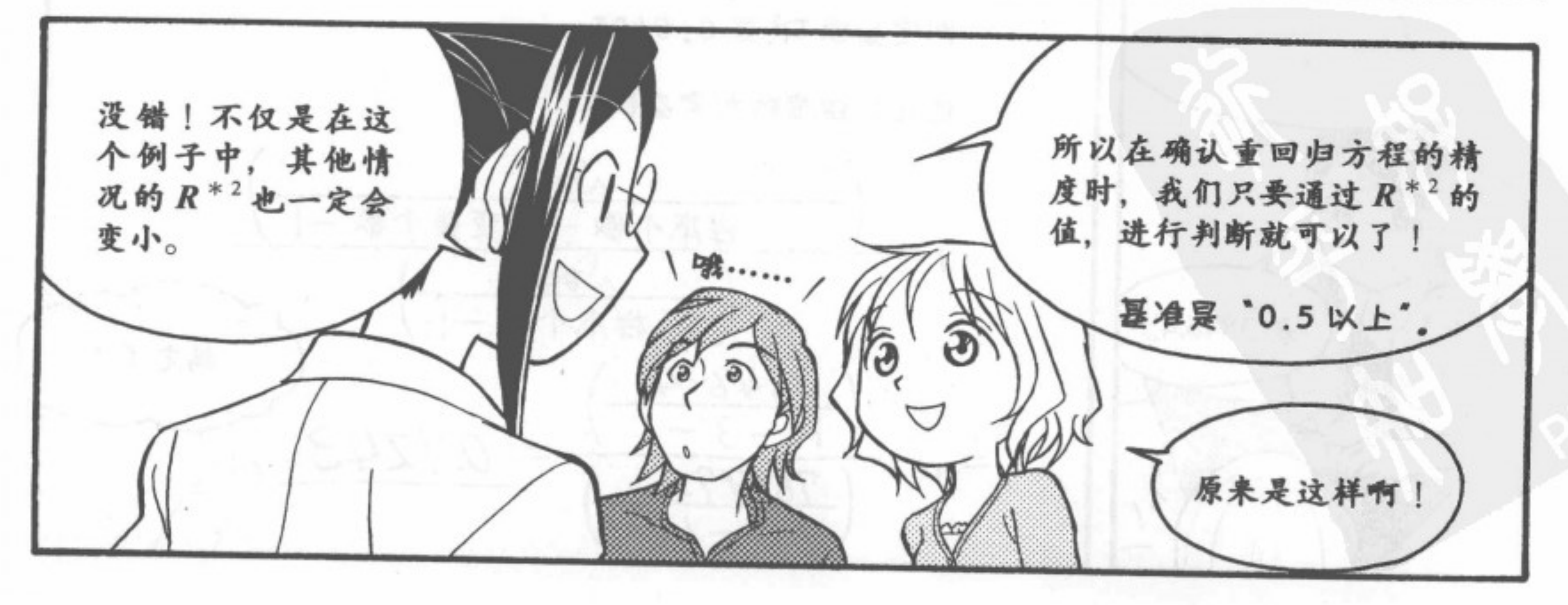
$$=1-\frac{\left(\dfrac{3846.4}{10-3-1}\right)}{\left(\dfrac{76199.6}{10-1}\right)}=0.9243$$

搞定了！


哈哈！
修正自由度的判定系数 R^{*2} 的值在不加"店长年龄"的情况①下会比较大啊！
① "店铺面积" "距最近车站的距离"
② "店铺面积"和 "距最近车站的距离" "店长年龄"
R^2 0.9452 < 0.9495
R^{*2} 0.9296 > 0.9243
真的啊！
怎么样？同 R^2 相比，还是 R^{*2} 的值更有说服力吧！


哎呀？

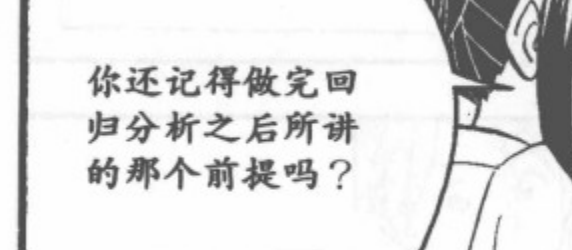

仔细看看，在①和②中，同判定系数 R^2 相比，修正自由度的判定系数 R^{*2} 的值都变小了呀！
① "店铺面积" "距最近车站的距离"
② "店铺面积"和 "距最近车站的距离" "店长年龄"
R^2 0.9452 0.9495
R^{*2} 0.9296 0.9243


没错！不仅是在这个例子中，其他情况的 R^{*2} 也一定会变小。
所以在确认重回归方程的精度时，我们只要通过 R^{*2} 的值，进行判断就可以了！
基准是"0.5 以上"。
哦……
原来是这样啊！


喂！


重回归分析也是需要推测总体情况的。


是“回归系数的检验”和总体回归的估计吗？
在重回归分析的情况下，称为“偏回归系数的检验”。


你还记得做完回归分析之后所讲的那个前提吗？


ぐっ
嗯……
那个在任何时候都一定成立的假设。


没错！
重回归分析也是如此！
~假设~


~假设~
“店铺面积为 x_1 坪，
距最近车站距离为 x_2m 时
月营业额”服从
平均值为 $A_1x_1+A_2x_2+B$，标准差为 σ 的
正态分布。
只有确信这个假设一定成立之后，才能进行分析！
是的！

④ 进行“回归系数的检验”

证明过程就不讲了，不过，说到 A_1、A_2、B、σ，可是统计学中常常会用到的，要记住哟！



所求出的重回归方程

$$y = a_1x_1 + a_2x_2 + b$$

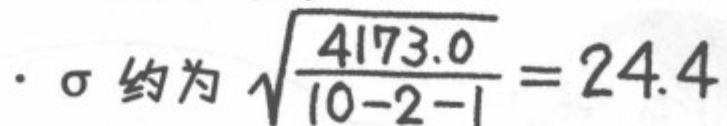
其中

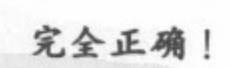
- A_1 约为 a_1
- A_2 约为 a_2
- B 约为 b
- σ 约为 $\sqrt{\dfrac{S_e}{\text{样本个数}-\text{自变量个数}-1}}$



是这样的吧？

- A_1 约为 41.5
- A_2 约为 −0.3
- B 约为 65.3
- σ 约为 $\sqrt{\dfrac{4173.0}{10-2-1}} = 24.4$









一种是"全面讨论偏回归系数的检验"，

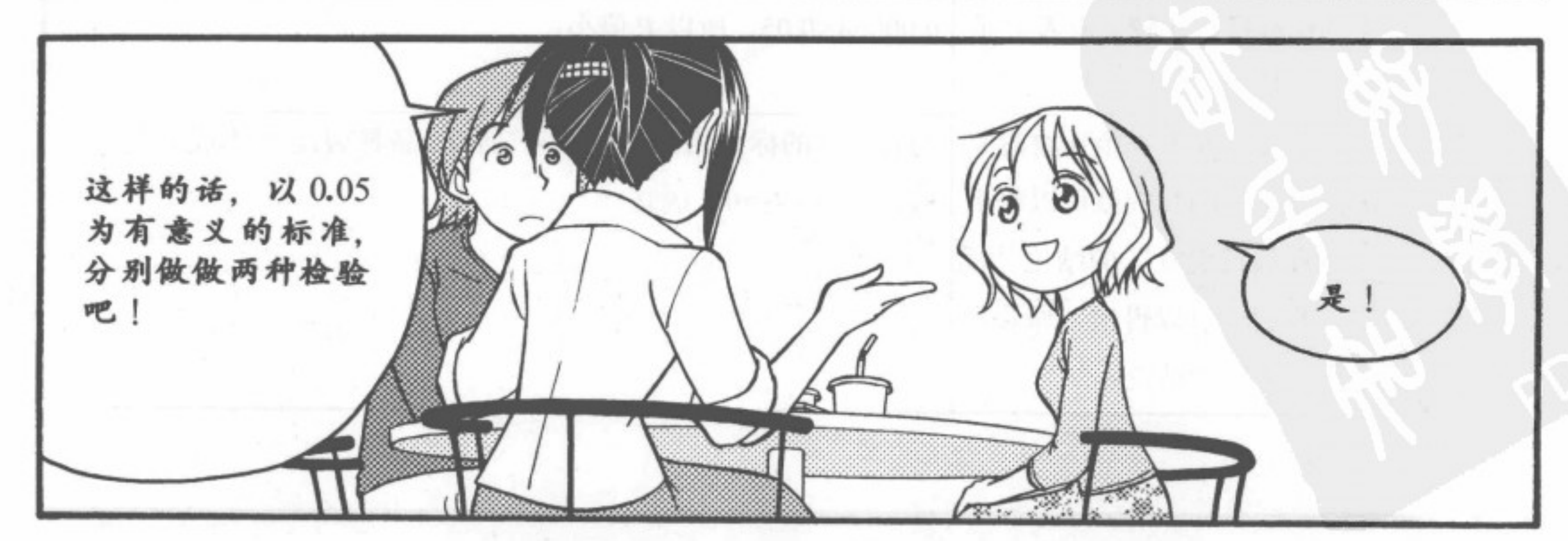
原假设	$A_1 = A_2 = 0$
备择假设	$A_1 = A_2 = 0$ 不成立。 即，下面任意一组关系成立： $A_1 \neq 0$ 且 $A_2 \neq 0$ $A_1 \neq 0$ 且 $A_2 = 0$ $A_1 = 0$ 且 $A_2 \neq 0$

另一种是"分别讨论偏回归系数的检验"。

原假设	$A_1 = 0$
备择假设	$A_1 \neq 0$







首先，进行“全面讨论偏回归系数的检验”！



步骤 1	定义总体。	将“店铺面积 x_1 坪、距最近车站的距离 $x_2 m$ 的店铺”作为总体。
步骤 2	建立原假设和备择假设。	原假设为“$A_1=A_2=0$ 成立”。 备择假设为“$A_1=A_2=0$ 不成立”。
步骤 3	选择所要进行的“检验”类型。	进行“全面讨论偏回归系数的检验”。
步骤 4	设定有意义的标准。	以 0.05 为有意义的标准。
步骤 5	通过样本数据求出检验统计量的值。	下面进行“全面讨论偏回归系数的检验”“全面讨论偏回归系数的检验”的检验统计量为 $\frac{S_{yy}-S_e}{\text{自变量的个数}} \div \frac{S_e}{\text{样本个数}-\text{自变量的个数}-1}$ 所以在本例题中的检验统计量的值为 $\frac{76199.6-4173.0}{2} \div \frac{4173.0}{10-2-1}=60.4$ 在本例题中，如果原假设成立，那么检验统计量就服从第 1 自由度为 2（= 自变量的个数）、第 2 自由度为 7（= 样本个数－自变量个数－ 2）的 F 分布。
步骤 6	再将步骤 5 中求出的检验统计量的值所对应的 P 值，与有意义的标准进行比较，看看 P 值是否比其小。	有意义的标准是 0.05。检验统计量的值为 60.4，所以 P 值为 0.00004。 $0.00004<0.05$，所以 P 值小。
步骤 7	如果在步骤 6 中 P 值比有意义的标准小，则我们就可以得出“备择假设成立”的结论。反之，我们就可以得出“原假设并没有错”的结论。	与有意义的标准相比，P 值小。所以，备择假设“不成立”，假设“$A_1=A_2=0$”成立。

接下来，进行“分别讨论偏回归系数的检验”！
我们以 A_1 为检验对象，来示范一下！



步骤 1	定义总体。	将“店铺面积 x_1 坪、距最近车站的距离 x_2m 的店铺”作为总体。
步骤 2	建立原假设和备择假设。	原假设为“ $A_1 = 0$ 成立”。 备择假设为“ $A_1 \neq 0$ 成立”。
步骤 3	选择所要进行的“检验”类型。	进行“分别讨论偏回归系数的检验”。
步骤 4	设定有意义的标准。	以 0.05 为有意义的标准。
步骤 5	通过样本数据求出检验统计量的值。	下面进行“分别讨论偏回归系数的检验”的过程。 “分别讨论偏回归系数的检验”的检验统计量为 $\frac{a_1^2}{S^{11※}} \div \frac{S_e}{样本个数-自变量的个数-1}$ 所以在本例题中的检验统计量的值为 $\frac{41.5^2}{0.0657} \div \frac{4173.0}{10-2-1} = 44.0$ 。 在本例题中，如果原假设成立，那么检验统计量就服从第 1 自由度为 1、第 2 自由度为 7（= 样本个数 − 自变量个数 − 2）的 F 分布。
步骤 6	再将步骤 5 中求出的检验统计量的值所对应的 P 值，与有意义的标准进行比较，看看 P 值是否比其小。	有意义的标准是 0.05。检验统计量的值为 44.0，所以 P 值为 0.0003。0.0003<0.05，所以 P 值小。
步骤 7	如果在步骤 6 中 P 值比有意义的标准小，则我们就可以得出“备择假设成立”的结论。反之，我们就可以得出“原假设成立”结论。	与有意义的标准相比，P 值小。所以，备择假设“$A_1 \neq 0$”成立。

※ S^{11} 的求解方法见下页说明。

但是，不管步骤7的结论是什么，我们还是习惯性地认为“只有在检验统计量

$$\frac{a_1^2}{S^{11}} \div \frac{S_e}{样本个数-自变量个数-1}$$

的值大于2的情况下，这个偏回归系数所对应自变量才对因变量的预测有意义。”



这是步骤5中出现的“S^{11}”。

$$\left[\begin{pmatrix} 10 & 8 & 8 & 5 & 7 & 8 & 7 & 9 & 6 & 9 \\ 80 & 0 & 200 & 200 & 300 & 230 & 40 & 0 & 330 & 180 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 10 & 80 & 1 \\ 8 & 0 & 1 \\ 8 & 200 & 1 \\ 5 & 200 & 1 \\ 7 & 300 & 1 \\ 8 & 230 & 1 \\ 7 & 40 & 1 \\ 9 & 0 & 1 \\ 6 & 330 & 1 \\ 9 & 180 & 1 \end{pmatrix}\right]^{-1} = \begin{pmatrix} 0.0657 & \cdots & \cdots \\ \cdots & 0.00001 & \cdots \\ \cdots & \cdots & \cdots \end{pmatrix}$$

也就是说，这是“S^{22}”。

店铺面积

距最近车站的距离

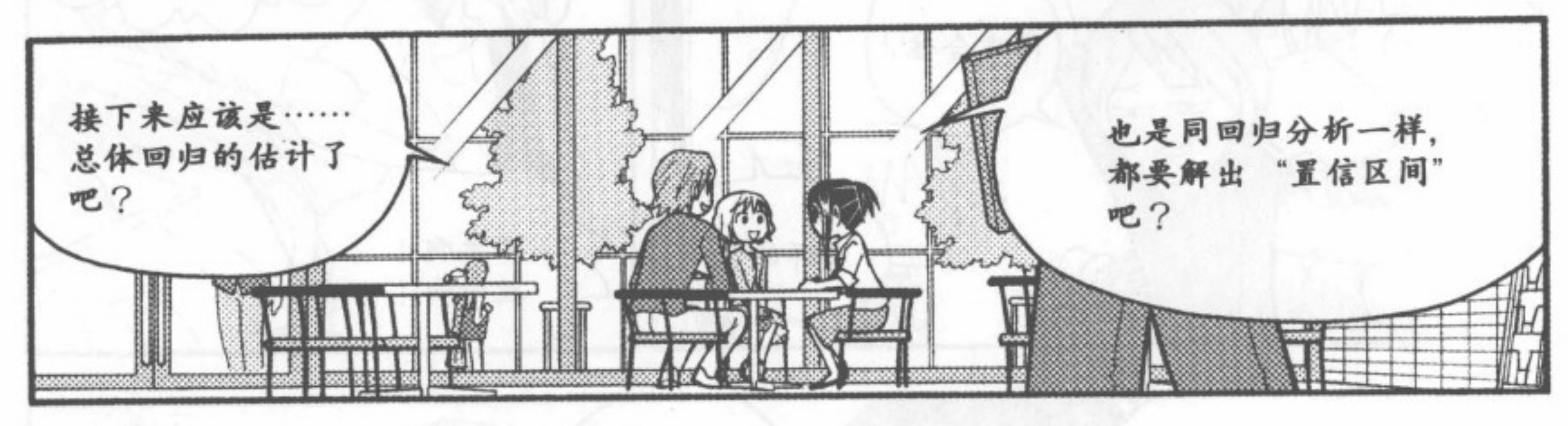
为了便于计算必须将这里都写成1



有些参考资料中，并不是依据F分布，而是依据t分布来讲解“偏回归系数的检验”的。
这个问题从数学的角度解释起来比较困难，所以我们不做详细介绍。但是，无论依据哪种分布，其最终的结论都是相同的。



⑤ 总体回归 $A_1x_1 + A_2x_2 + \cdots + A_px_p + B$ 的估计







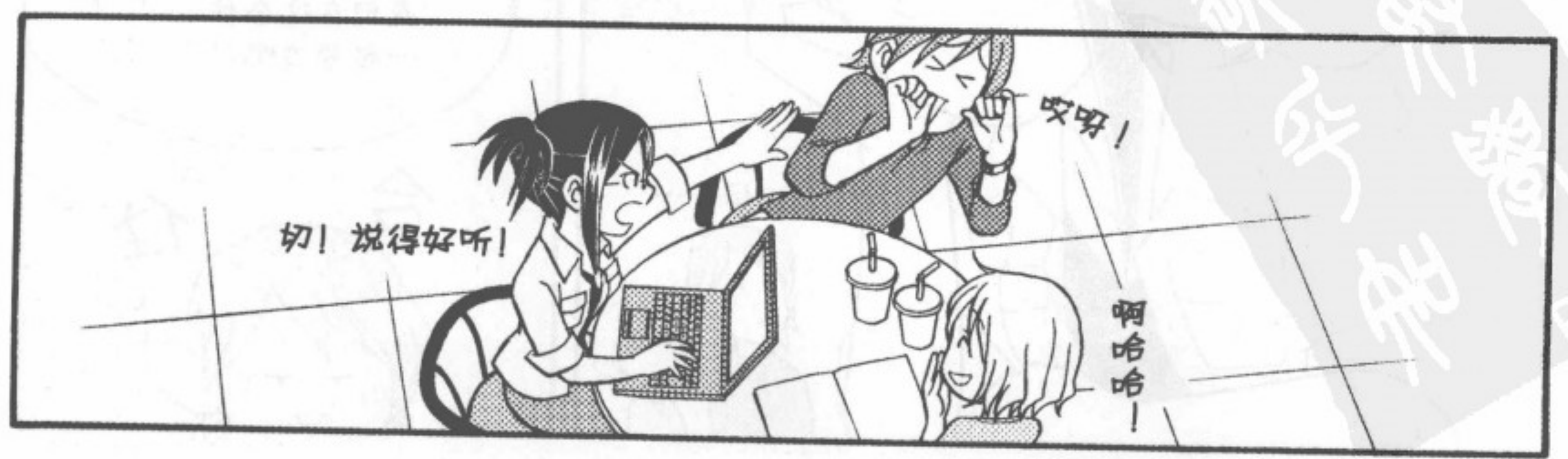


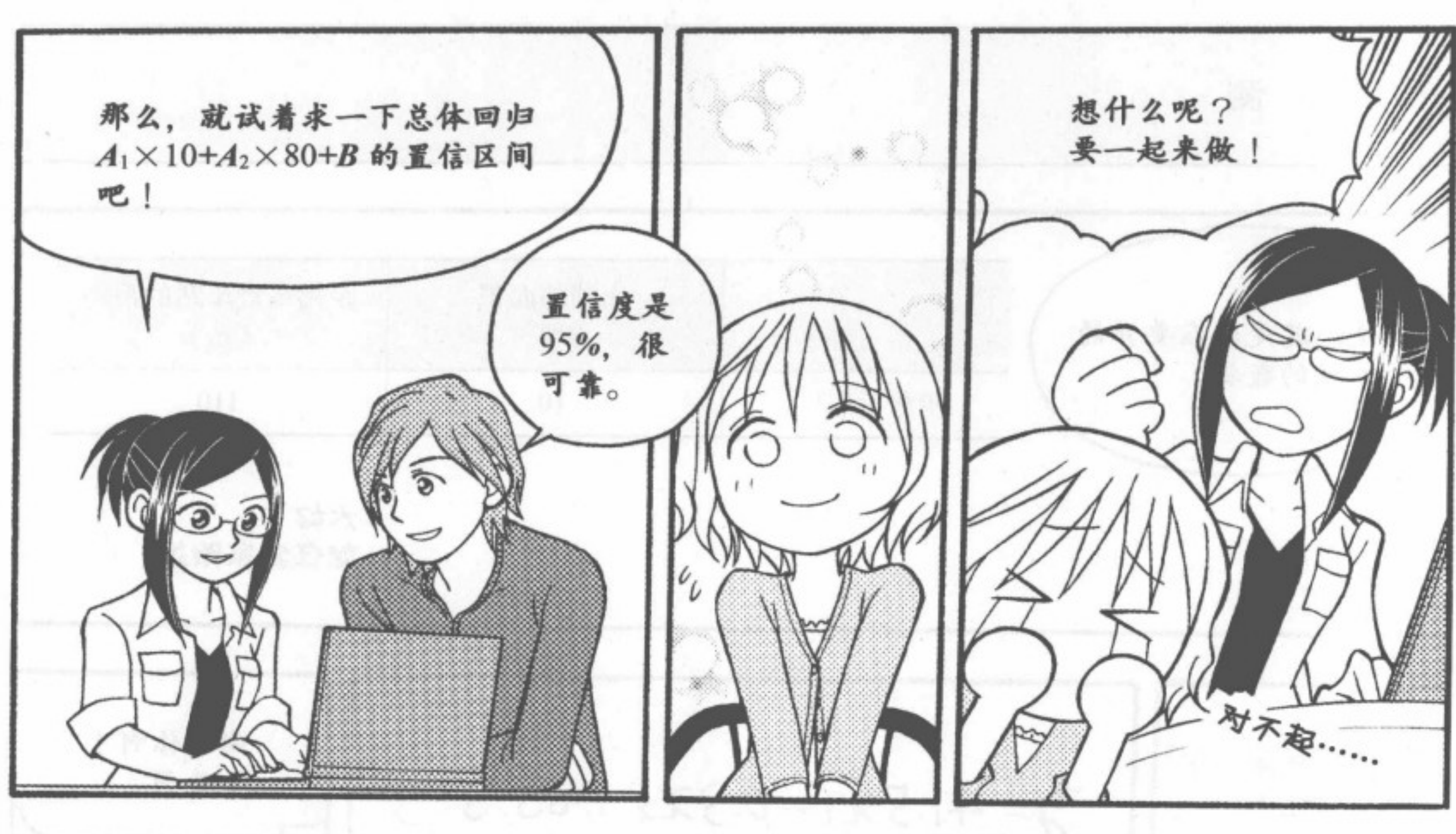



睡觉可不
行哦！
好！
枕头吗不可以扔枕头哦……
同意合住！
那样的话？
那么，就通宵
计算吧！
不要啊！
救命啊！


那现在，我们就用数据
分析软件来求置信区间
吧！
电脑真是太
伟大啦！


虽然很可靠，不过
我们也不能偷懒。
这个嘛！
只有在关键的
时刻才可以使
用哦！


哎呀！
切！说得好听！
啊哈哈！


那么，就试着求一下总体回归
$A_1 \times 10 + A_2 \times 80 + B$ 的置信区间
吧！
置信度是
95%，很
可靠。
想什么呢？
要一起来做！
对不起……


453.2 加减 34.9。
※求解方法请参见138页。


那具体应该是
……
453.2+34.9 和
453.2-34.9……


总体回归 $A_1 \times 10 + A_2 \times 80 + B$ 在置
信度是 95% 的情况下，应该处于
418.3 万日元以上 488.1 万日元以
下，是这样吧？
算的没错！

⑥预　测

	店铺的面积（坪）	距离最近车站的距离（m）
伊势桥店	10	110



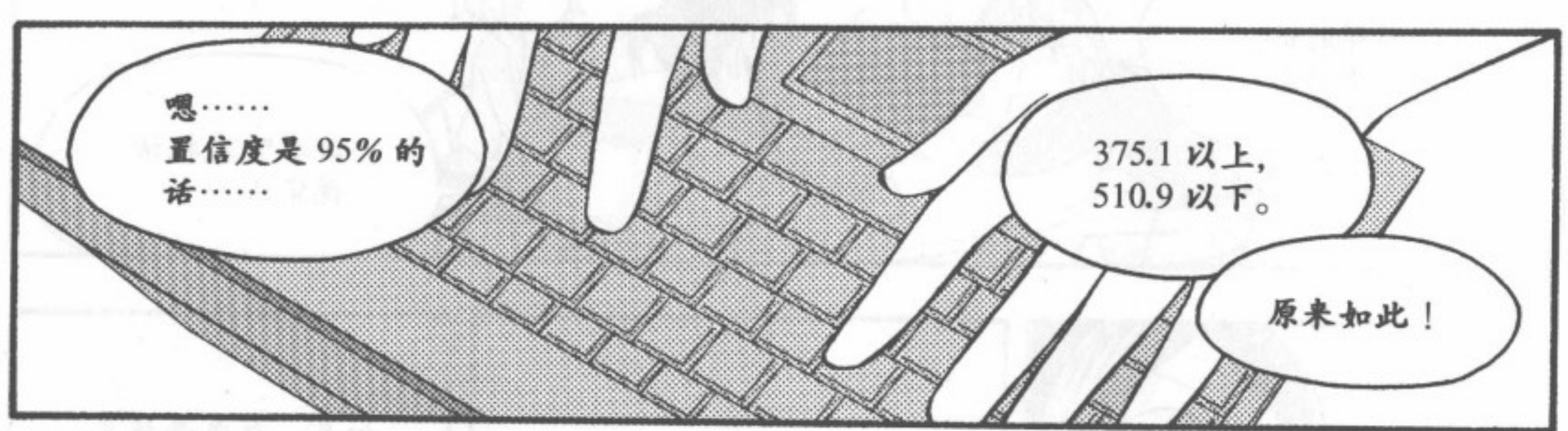
$$y = 41.5x_1 - 0.3x_2 + 65.3$$
$$= 41.5 \times 10 - 0.3 \times 110 + 65.3$$
$$= 447.3$$

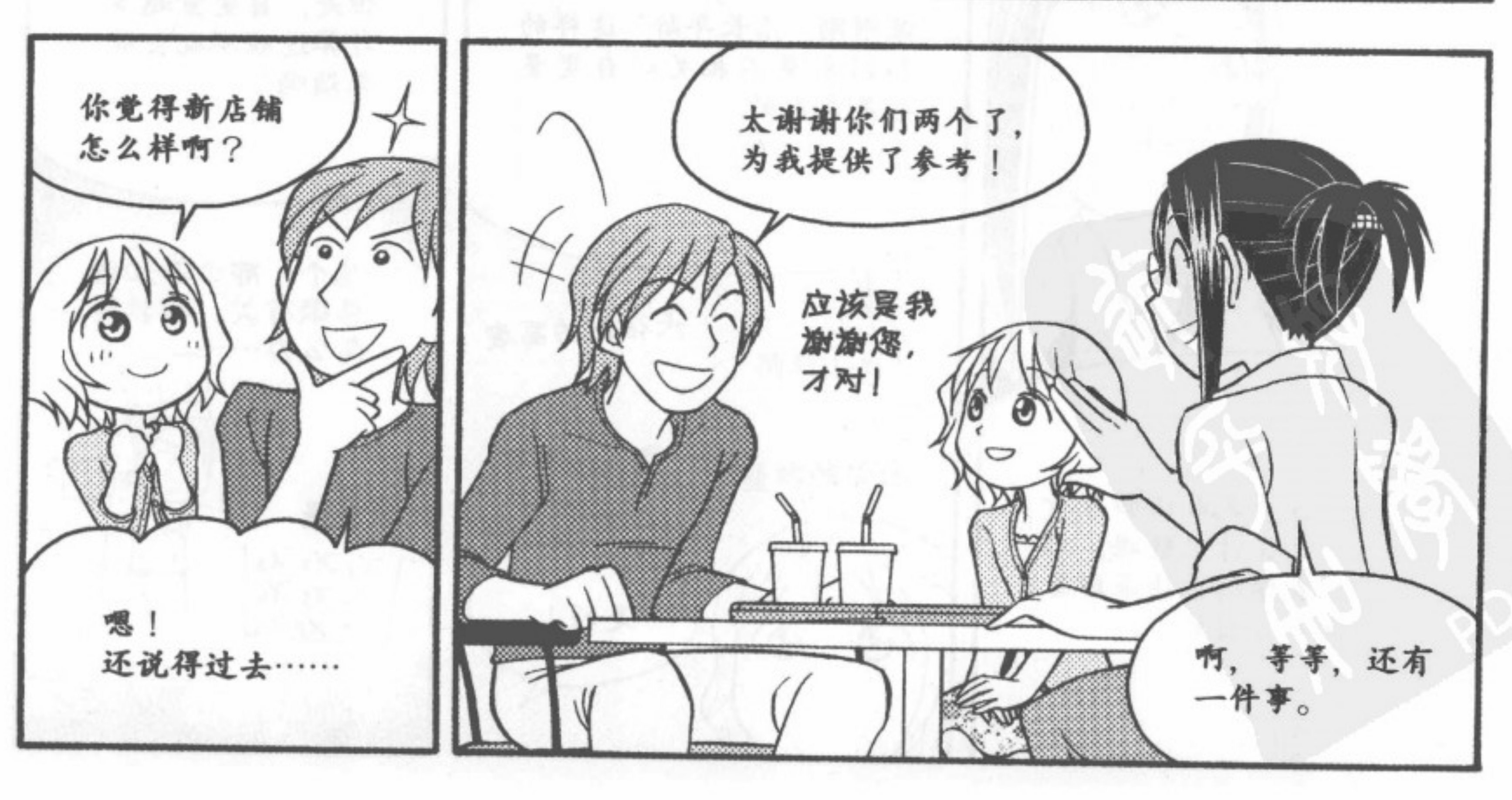



可是，那个"马…
马…"是什么呢？
什么，什么
距离的……
プッ


这个！
在求解预测区间
的时候，已经用
过马氏距离。
你们两个要记住
名字啊……


那，这个也是靠
软件计算的吧！
合作真让人为难啊！
是啊！
是啊！


嗯……
置信度是 95% 的
话……
375.1 以上，
510.9 以下。
原来如此！


你觉得新店铺
怎么样啊？
嗯！
还说得过去……
太谢谢你们两个了，
为我提供了参考！
应该是我
谢谢您，
才对！
啊，等等，还有
一件事。


最后作为赠品，再介绍一种更好的求解重回归方程的方法。


啊啊啊？

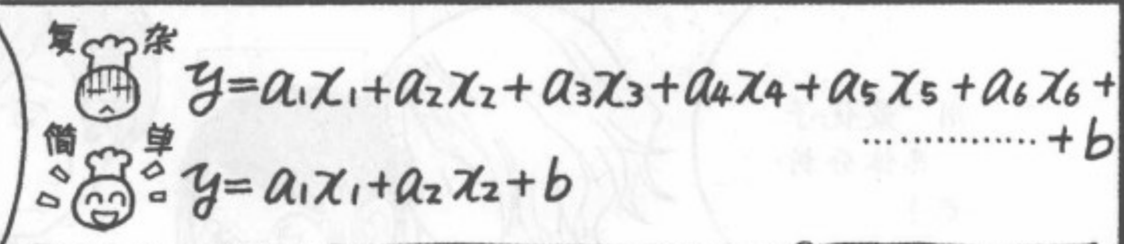

如果是那样的话，为什么最开始不……
认真听我把话说完！


重回归分析也和回归分析一样，无论什么数据，都是通过数学的方法求解基本重回归方程的。


像刚刚“店长年龄”这样的和因变量不相关的自变量也是存在的。
店长年龄
天花板的高度
座位的数量
餐盘的数量
是的！


但是，自变量越多计算过程不就会越麻烦吗？
这个，那个都和销售额有关，那我要怎么办……
自变量
x_1 x_2 x_3
x_4 x_5 x_6
x_7 x_8 x_9

考虑到以上情况，对于分析者来说，“更好的重回归方程”可以说是“自变量个数不多、并且精度又高的重回归方程”。



精度高 $y=a_1x_1+a_2x_2+b$ R^{*2} > $y=a_1x_1+a_3x_3+b$ R^{*2} 精度低



是这样啊！

这种“自变量个数不多、并且精度高的重回归方程”的求解方法是：



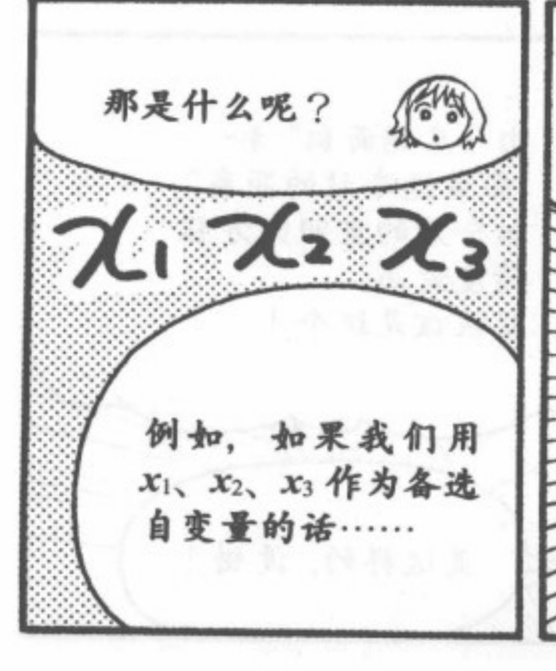
- 变量增加法
- 变量减少法
- 变量增减法
- 基于“情报量标准”的方法

除此之外，还有许多种类的求解方法，不过……

今天我就来讲一种比这些方法都简洁易懂的方法——“最优子集法”。

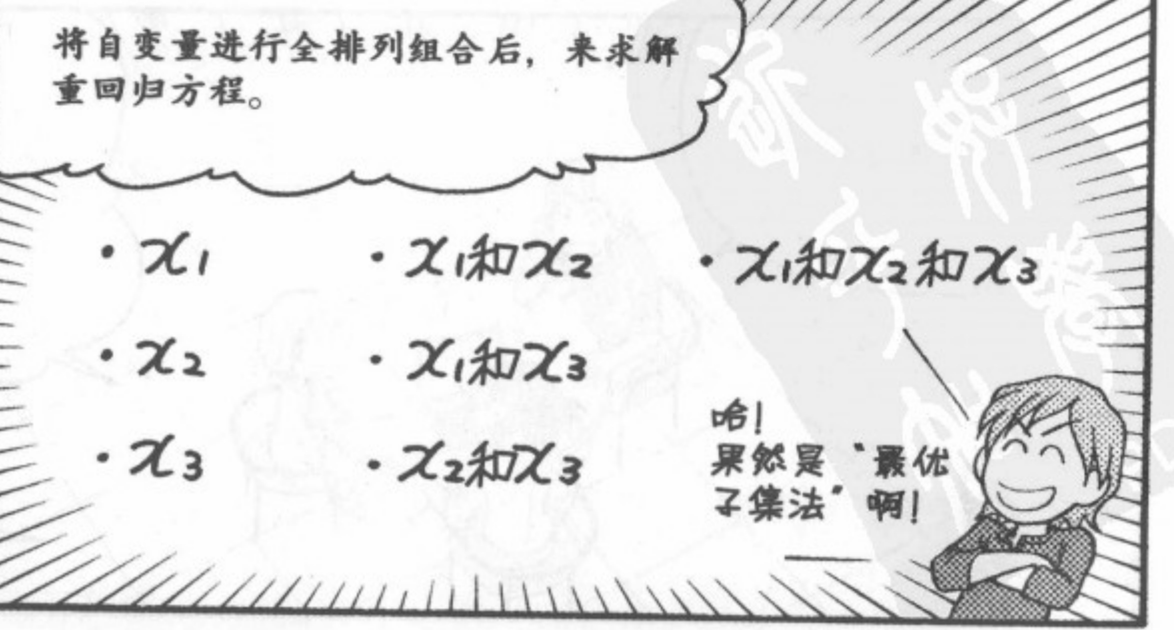


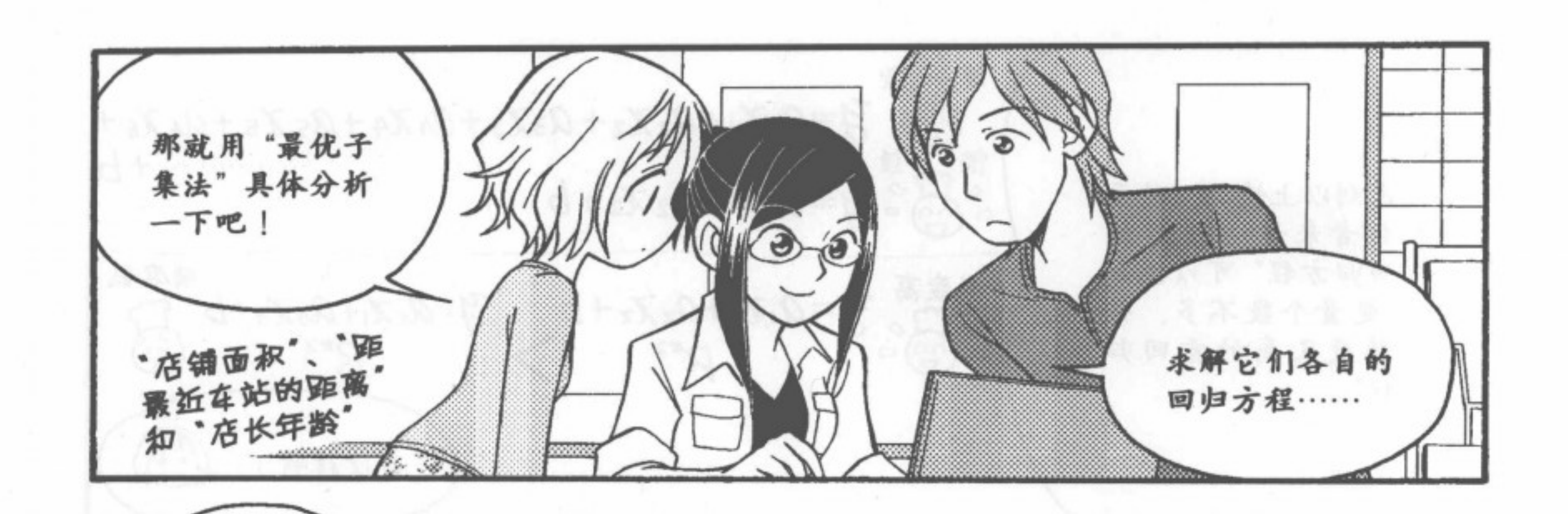
那是什么呢？



例如，如果我们用 x_1、x_2、x_3 作为备选自变量的话……

将自变量进行全排列组合后，来求解重回归方程。



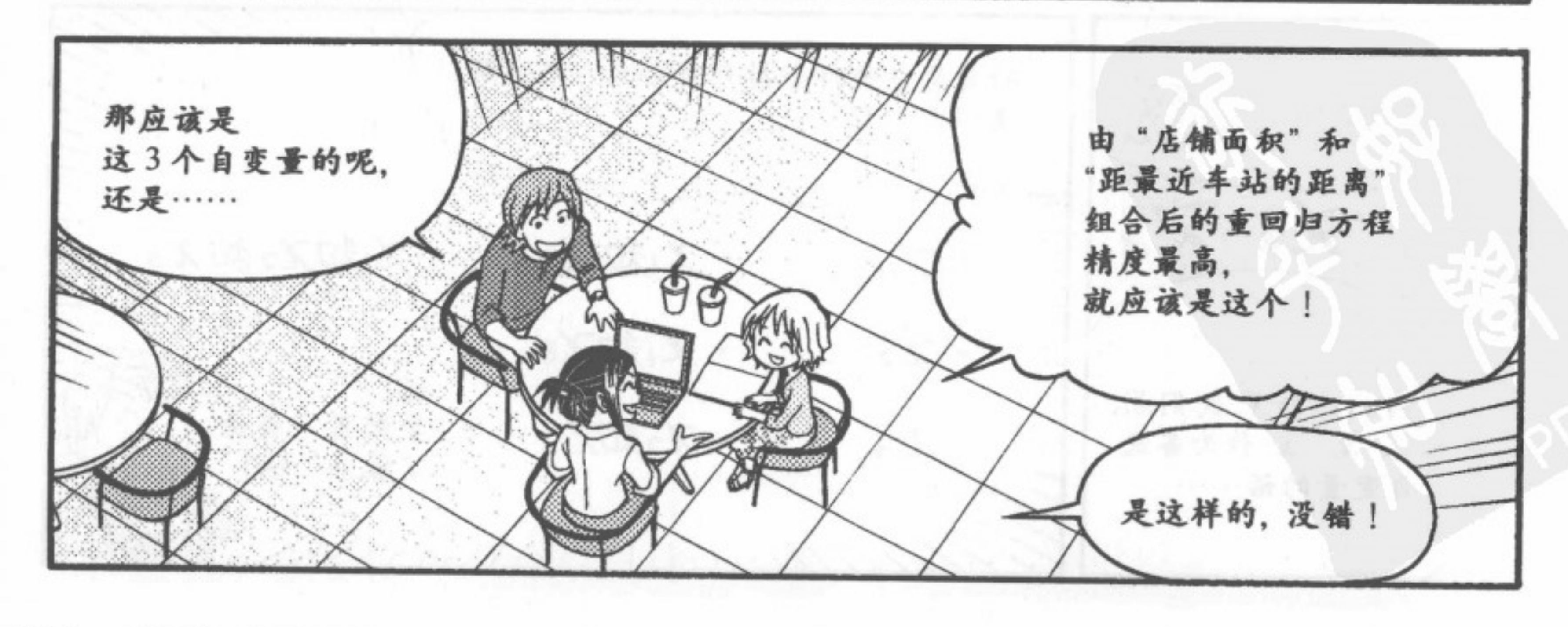


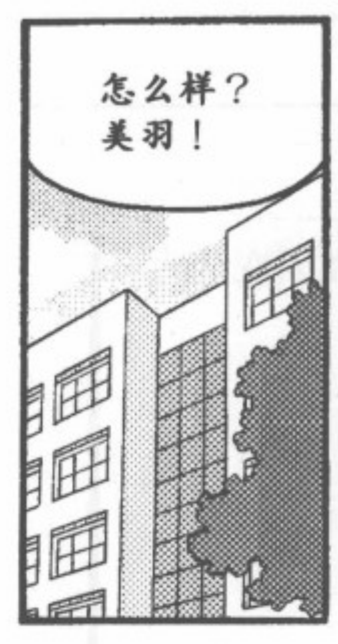
就是这样了。

哈哈！

	a_1	a_2	a_3	b	修正自由度的判定系数R^{*2}
1	54.9			−91.3	0.7709
2		−0.6		424.8	0.5508
3			0.6	309.1	0.0000
1和2	41.5	−0.3		65.3	0.9296
1和3	55.6		2.0	−170.1	0.7563
2和3		−0.6	−0.4	438.9	0.4873
1和2和3	42.2	−0.3	1.1	17.7	0.9243

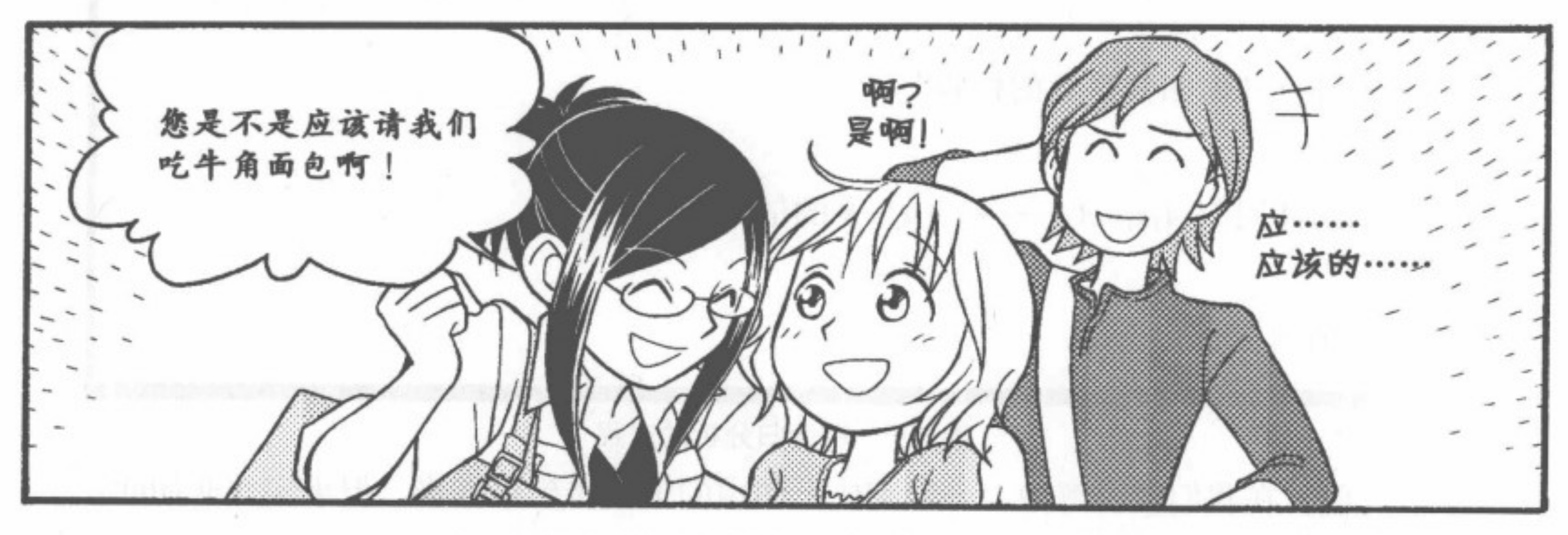
自变量为 1 和 2，也就是“店铺面积”和“距最近车站的距离”的重回归方程是 $y=41.5x_1-0.3x_2+65.3$。




怎么样？
美羽！


能理解重回归分析了吗？
是啊！
真是谢谢您啦！
我也学会了。


那么，未来的店长先生，
作为数据分析的费用……
チラ…


您是不是应该请我们吃牛角面包啊！
啊？
是啊！
应……
应该的……


噢，回去做冰红茶喽！
？
小美羽！
小美羽！


那家伙，是不是很不错啊？




是啊，非常不错！
你们两个，快点！

✿3. 重回归分析过程中的注意事项✿

下图中，我们再次给出 106 页中出现的重回归分析的过程。

① 首先，为了讨论是否具有求解回归方程的意义，画出各自变量和因变量的散点图。

↓

② 求解重回归方程。

↓

③ 确认重回归方程的精度。

↓

④ 进行“偏回归系数的检验”。

↓

⑤ 总体回归 $A_1x_1+A_2x_2+\cdots+A_px_p+B$ 的估计。

↓

⑥ 预测。

图 3.1　重回归分析的过程

此前，在我们的讲解中，必须完成上图中的第①步到第⑤步。但事实并非如此。同回归分析一样，不同的情况下，只完成第①步到第③步亦可。

但是在本章中提到的风见面包店只有 10 个店铺而已，而且举例中所用的店铺面积为 10 坪、距最近车站距离为 80m 的店铺，也只有“梦之丘总店”这一家。这样一来，在估计总体回归 $A_1\times10+A_2\times80+B$，或是在进行“偏回归系数检验”时，就可能会使读者产生疑问了。这也是情有可原的。实际上，理纱是在如下解释的基础上进行分析的。

“店铺面积为 10 坪、距最近车站距离为 80m”——像这样的店铺，风见面包店今后还会开设很多。这次只不过是从这样的店铺群中，随机抽取到“梦之丘总店”而已。

对于理纱的解释可以说还是有值得商榷的地方。笔者认为这是一个比较牵强的解释。严格地说，如果我们考虑到风见面包店的人气问题，那么就可以说它是一个“非常没有根据的解释”。既然如此，是否还需要专门进行总体回归的估计或是“检验”呢？笔者认为，还是要从记述统计学的角度进行分析，判断是否需要。

✿4. 标准化残差✿

同回归分析一样，这里也有必要对重回归分析的**标准化残差**进行讨论。重回归分析的标准化残差为

$$\frac{\text{残差}}{\sqrt{\dfrac{\text{残差平方和}}{\text{样本个数}-\text{自变量个数}-1}}}=\frac{y-\hat{y}}{\sqrt{\dfrac{S_e}{\text{样本个数}-\text{自变量个数}-1}}}$$

下表中记录的是本章例子中的标准化残差。

◆表 3.1　本章例子中的标准化残差

	店铺的面积 x_1	距最近车站的距离 x_2	月营业额 y	月营业额 $=41.5x_1-0.3x_2+65.3$	残差 $y-$	标准化残差 $\dfrac{y-}{\sqrt{\dfrac{4173.0}{10-2-1}}}$
梦之丘总店	10	80	469	453.2	15.8	0.6
寺井站大厦店	8	0	366	397.4	−31.4	−1.3
曾根店	8	200	371	329.3	41.7	1.7
桥本大街店	5	200	208	204.7	3.3	0.1
桔梗町店	7	300	246	253.7	−7.7	−0.3
邮政局前店	8	230	297	319.0	−22.0	−0.9
水道町站前店	7	40	363	342.3	20.7	0.8
六条站大厦店	9	0	436	438.9	−2.9	−0.1
若叶川店	6	330	198	201.9	−3.9	−0.2
美里店	9	180	364	377.6	−13.6	−0.6

$$\frac{-13.6}{\sqrt{\dfrac{4173.0}{10-2-1}}}=-0.6$$

标准化残差的绝对值大的个体，被看成与其他的个体性质不同。当绝对值大于3 的个体存在时，将其剔除之后，再进行重回归分析。

✿5. 马氏距离以及重回归分析中的置信区间和预测区间✿

如 127 页和 131 页所述，在计算重回归分析的置信区间和预测区间的过程中，出现了所谓的马氏距离（*Mahalanobis Distance*）。这与我们在初中和高中学过的普通距离——欧氏距离（*Euclidean Distance*）有所不同，是一种重新定义的距离概念。

读者可能会问“为什么要专门定义出这样一种距离呢？”对于这个问题，笔者本想进行回答，但是由于篇幅所限，并且与本书的写作主旨没有太大关系，所以只好点到即止。但是，稍后会讲解其计算方法。可是不管怎样，马氏距离在统计学中是赫赫有名的，所以无论何时都请记住它。顺便提一下，马氏“*Mahalanobis*”是数学家 *Prasanta Chandra Mahalanobis* 的名字。

接下来，我们讨论本节的主题。重回归分析的置信区间的求解顺序如下所述。为了方便起见，我们仍然使用第 129 页提到的“梦之丘总店”的例子，讲解其置信区间的求解过程。

步骤 1

$$\text{求}\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}\text{的逆矩阵}\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}^{-1} = \begin{pmatrix} S^{11} & S^{12} & \cdots & S^{1p} \\ S^{21} & S^{22} & \cdots & S^{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S^{p1} & S^{p2} & \cdots & S^{pp} \end{pmatrix}$$

例如，其中的“S_{22}”表示“第 2 个自变量的离差平方和”，“S_{25}”就表示“第 2 个自变量和第 5 个自变量的离差积和”，不难理解它与“S_{52}”是相等的。

$$\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}^{-1} = \begin{pmatrix} S^{11} & S^{12} \\ S^{21} & S^{22} \end{pmatrix} = \begin{pmatrix} 20.1 & -792 \\ -792 & 128840 \end{pmatrix}^{-1} = \begin{pmatrix} 0.0657 & 0.0004 \\ 0.0004 & 0.00001 \end{pmatrix}$$

这里所求得的 S^{11} 和 S^{22} 的值同第 126 页出现的值是相同的。同时不仅限于此例，

- $\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}^{-1}$ 中S^{ii}和S^{ij}的值
- “分别讨论偏回归系数的检验”中所求得的S^{ii}和S^{ij}的值，
  都一定是相同的。

步骤 2

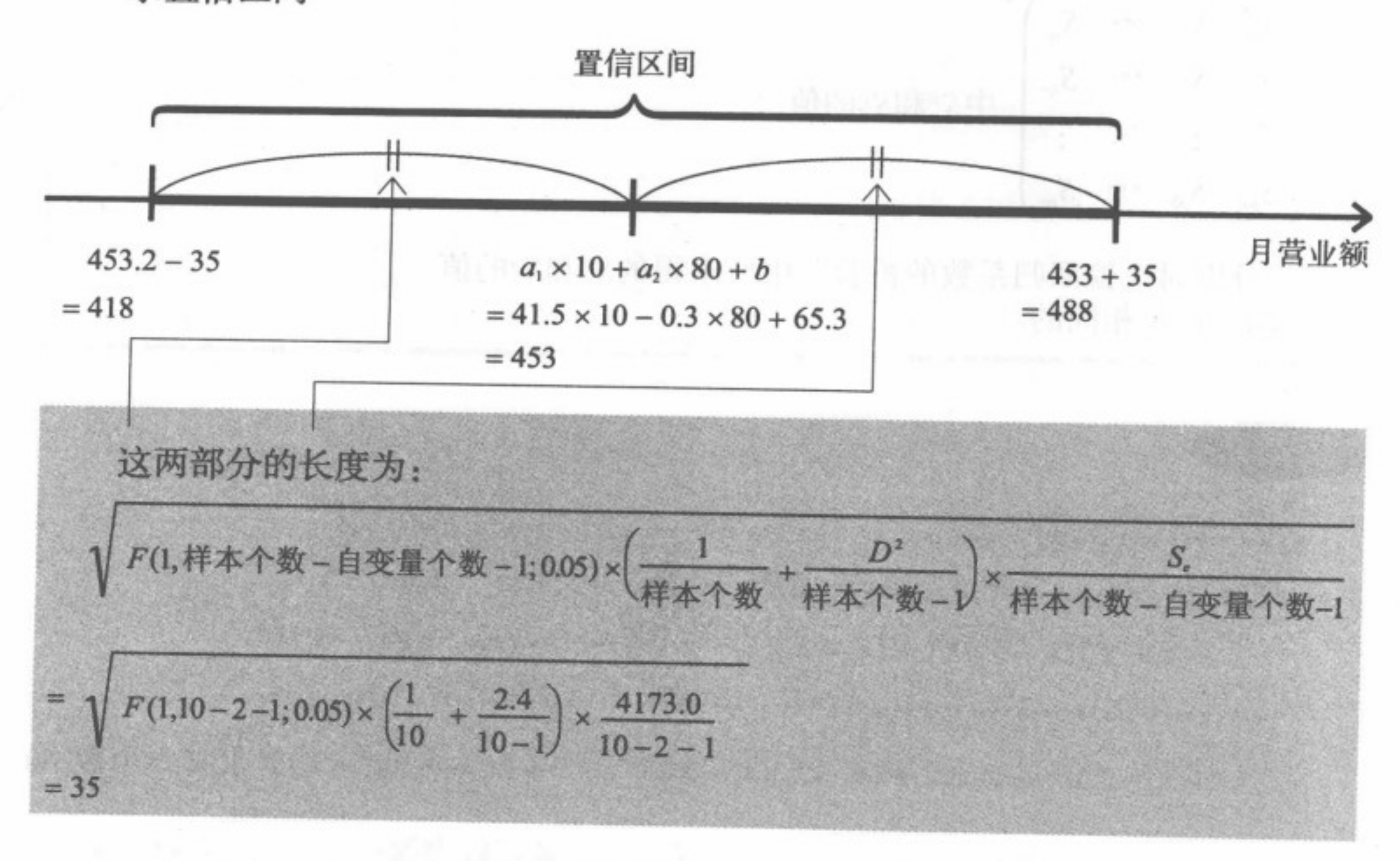
求“马氏距离的平方”。

$$\begin{aligned} D^2 = [&(x_1-\bar{x}_1)(x_1-\bar{x}_1)S^{11} + (x_1-\bar{x}_1)(x_2-\bar{x}_2)S^{12} + \cdots + (x_1-\bar{x}_1)(x_p-\bar{x}_p)S^{1p} \\ &+(x_2-\bar{x}_2)(x_1-\bar{x}_1)S^{21} + (x_2-\bar{x}_2)(x_2-\bar{x}_2)S^{22} + \cdots + (x_2-\bar{x}_2)(x_p-\bar{x}_p)S^{2p} \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &+(x_p-\bar{x}_p)(x_1-\bar{x}_1)S^{p1} + (x_p-\bar{x}_p)(x_2-\bar{x}_2)S^{p2} + \cdots + (x_p-\bar{x}_p)(x_p-\bar{x}_p)S^{pp}](\text{样本个数}-1) \end{aligned}$$

$$\begin{aligned} D^2 &= [(x_1-\bar{x}_1)(x_1-\bar{x}_1)S^{11} + (x_1-\bar{x}_1)(x_2-\bar{x}_2)S^{12} \\ &\quad + (x_2-\bar{x}_2)(x_1-\bar{x}_1)S^{21} + (x_2-\bar{x}_2)(x_2-\bar{x}_2)S^{22}](\text{样本的个数}-1) \\ &= [(10-7.7)(10-7.7)\times 0.0657 + (10-7.7)(80-156)\times 0.0004 \\ &\quad + (80-156)(10-7.7)\times 0.0004 + (80-156)(80-156)\times 0.00001](10-1) \\ &= 2.4 \end{aligned}$$

步骤 3

求置信区间



这两部分的长度为：

$$\sqrt{F(1,\text{样本个数}-\text{自变量个数}-1;0.05)\times\left(\frac{1}{\text{样本个数}}+\frac{D^2}{\text{样本个数}-1}\right)\times\frac{S_e}{\text{样本个数}-\text{自变量个数}-1}}$$

$$=\sqrt{F(1,10-2-1;0.05)\times\left(\frac{1}{10}+\frac{2.4}{10-1}\right)\times\frac{4173.0}{10-2-1}}$$

$= 35$

在求解预测区间的时候，同回归分析一样，区间宽度不是

$$\sqrt{F(1,\text{样本个数}-\text{自变量个数}-1;0.05)\times\left(\frac{1}{\text{样本个数}}+\frac{D^2}{\text{样本个数}-1}\right)\times\frac{S_e}{\text{样本个数}-\text{自变量个数}-1}}$$

而是

$$\sqrt{F(1,\text{样本个数}-\text{自变量个数}-1;0.05)\times\left(1+\frac{1}{\text{样本个数}}+\frac{D^2}{\text{样本个数}-1}\right)\times\frac{S_e}{\text{样本个数}-\text{自变量个数}-1}}$$

当置信度是 99% 的时候，只需将

$F(1,\text{样本个数}-\text{自变量个数}-1;0.05)=F(1,10-2-1;0.05)=5.6$ 这一部分，

替换为

$F(1,\text{样本个数}-\text{自变量个数}-1;0.01)=F(1,10-2-1;0.01)=12.2$ 就可以了 。

✿ 6. 自变量为分类数据时的重回归分析 ✿

下面，我们再次给出第 107 页出现的表格。

◆表 3.2　第 107 页出现的表

	店铺面积 (坪)	距最近车站距离 (m)	月营业额 (万日元)
梦之丘总店	10	80	469
寺井站大厦店	8	0	366
曾根店	8	200	371
桥本大街店	5	200	208
桔梗町店	7	300	246
邮政局前店	8	230	297
水道町站前店	7	40	363
六条站大厦店	9	0	436
若叶川店	6	330	198
美里店	9	180	364

由上表可知，作为自变量的“店铺面积”和“距最近车站的距离”以及作为因变量的“月营业额”都是数值数据。

在重回归分析中，因变量必须为可测变量，而自变量则可以是

- 仅为数值数据
- 数值数据和分类数据的混合
- 仅为分类数据

这三种之一。

下面我们举两个数值数据和分类数据混合时的例子，和一个仅为分类数据的例子。

■ 数值数据和分类数据混合时的例子 <1>

	店铺面积(坪)	距最近车站距离(m)	有品尝专柜	无品尝专柜	月营业额(万日元)
梦之丘总店	10	80	1	0	469
寺井站大厦店	8	0	0	1	366
曾根店	8	200	1	0	371
桥本大街店	5	200	0	1	208
桔梗町店	7	300	0	1	246
邮政局前店	8	230	0	1	297
水道町站前店	7	40	0	1	363
六条站大厦店	9	0	1	0	436
若叶川店	6	330	0	1	198
美里店	9	180	1	0	364

"1"表示"符合(条件)"、 "0"表示"不符合(条件)"
如第47页所述,分析时,这2列中有必要省略1列。
这里我们省略"无品尝专柜"这一列。

那么,通过对以上数据的分析,我们可以求出其重回归方程为:

$$y = 30.6x_1 - 0.4x_2 + 39.5x_3 + 135.9$$

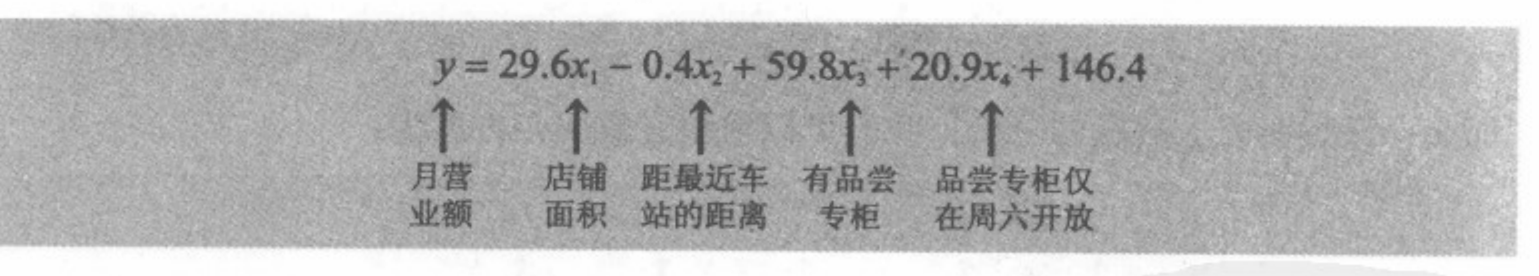
y:月营业额 x_1:店铺面积 x_2:距最近车站的距离 x_3:有品尝专柜

■ 数值数据和分类数据混合时的例子 <2>

	店铺面积(坪)	距最近车站距离(m)	有品尝专柜	品尝专柜仅在周六开放	无品尝专柜	月营业额(万日元)
梦之丘总店	10	80	1	0	0	469
寺井站大厦店	8	0	0	0	1	366
曾根店	8	200	1	0	0	371
桥本大街店	5	200	0	0	1	208
桔梗町店	7	300	0	0	1	246
邮政局前店	8	230	0	0	1	297
水道町站前店	7	40	0	0	1	363
六条站大厦店	9	0	0	1	0	436
若叶川店	6	330	0	0	1	198
美里店	9	180	0	1	0	364

"1"表示"符合（条件）"、"0"表示"不符合条件"。
如第 47 页所述，分析时，这 3 列中有必要省略 1 列。
这里我们省略"无有品尝专柜"这一列。

那么，通过对以上数据的分析，我们可以求出其重回归方程为：



■ 仅为分类数据的例子

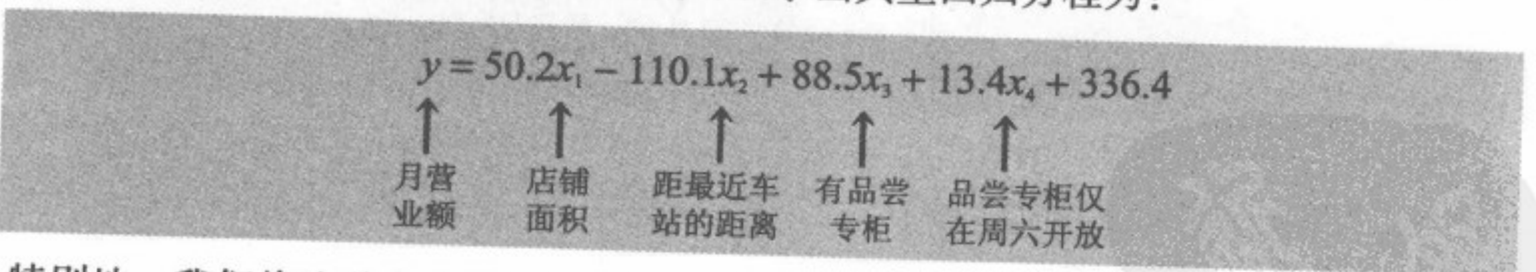
	店铺面积(坪)	距最近车站距离(m)	有品尝专柜	品尝专柜仅在周六开放	月营业额(万日元)
梦之丘总店	1	0	1	0	469
寺井站大厦店	1	0	0	0	366
曾根店	1	1	1	0	371
桥本大街店	0	1	0	0	208
桔梗町店	0	1	0	0	246
邮政局前店	1	1	0	0	297
水道町站前店	0	0	0	0	363
六条站大厦店	1	0	0	1	436
若叶川店	0	1	0	0	198
美里店	1	0	0	1	364

"1"表示"8 坪以上"。
"0"表示"不足 8 坪"。

"1"表示"200m 以上"。
"0"表示"不足 200m"。

"1"表示"符合(条件)"、 "0"表示"不符合条件"。

那么，通过对以上数据的分析，我们可以求出其重回归方程为：



特别地，我们将这种自变量仅为分类数据的重回归分析称为数量化 I 类。

✿7.多重共线性✿

本节的内容有些难，所以只做简单介绍，点到即止。

如果自变量之间存在很强的相关性，就会出现如下奇怪的情况：

• 求不出偏回归系数

• 即便可以求出偏回归系数，也会出现本应是正值的地方，不知道为什么会求出负值。

在数学上，当出现

• $\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}$ 的行列式[1]的值为 0。

• $\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{pmatrix}$ 的行列式的值近似为 0。

这样的情况时，我们就将其称为“存在多重共线性问题”。

我们可以通过 *VIF*[2] 或容许度[3]（Tolerance）这样的指标，来判断是否“存在多重共线性问题”。在 Excel 中，求行列式的值所使用的“MDETERM”函数，也是一种求行列式的值的方法。

无论在学习重回归分析时这个问题有多么深奥，读者只需掌握到以下程度即可，也就是“当自变量之间存在很强的相关性时，我们可以省去其中任意一个自变量后再进行分析”。

1. 本书不做说明。
2. 本书不做说明。
3. 本书不做说明。

✿ 8. “各自变量对因变量的影响”和重回归分析 ✿

对于在本书中初次接触重回归分析的读者，可以跳过以下内容不做阅读。

重回归分析，不仅可以用作一种预测手段，而且也可以用作一种调查“各自变量对因变量的影响”的方法。

我们来读下面这个故事。

鸟越先生是一家糖果公司的产品开发研究员。鸟越先生最近负责的糖果销量非常好。为了找出这种糖果畅销的原因，公司邀请了一些评论人员来试吃糖果。以下便是当时使用的调查问卷。

问题

您对这种糖果有何评价（每一项只能画一个○）

$Q1$ 味道	1 不喜欢　2 一般　3 喜欢
$Q2$ 分量	1 不喜欢　2 一般　3 喜欢
$Q3$ 便于食用	1 不喜欢　2 一般　3 喜欢
$Q4$ 包装设计	1 不喜欢　2 一般　3 喜欢
$Q5$ 综合满意度	1 不喜欢　2 一般　3 满意

下面的表格记录的是调查结果。

	Q1. 味道	Q2. 分量	Q3. 便于食用	Q4. 包装设计	Q5. 综合满意度
回答者1	2	2	3	2	2
回答者2	1	1	3	1	3
回答者3	2	2	1	1	1
回答者4	3	3	3	2	2
回答者5	1	1	2	2	1
回答者6	1	1	1	1	1
回答者7	3	3	1	3	3
回答者8	3	3	1	2	2
回答者9	3	3	1	2	3
回答者10	1	1	3	1	1
回答者11	2	3	2	1	3
回答者12	2	1	1	1	1
回答者13	3	3	3	1	3
回答者14	3	3	1	3	3
回答者15	3	2	1	1	2
回答者16	1	1	3	3	1
回答者17	2	2	2	1	1
回答者18	1	1	1	3	1
回答者19	3	1	3	3	3
回答者20	3	3	3	3	3

将变量逐一标准化[1]以后，对上表中的数据进行分析。于是可以推导出如下重回归方程：

$$y = 0.41x_1 + 0.32x_2 + 0.26x_3 + 0.11x_4$$

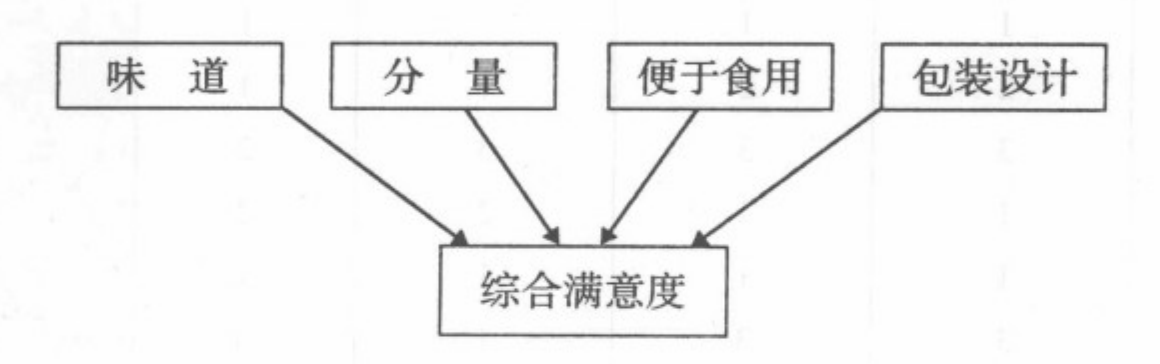
Q5. 综合满意度 (y)　Q1. 味道 (x_1)　Q2. 分量 (x_2)　Q3. 便于食用 (x_3)　Q4. 包装设计 (x_4)

观察偏回归系数[2]的值的大小，可知“*Q*1. 味道”的值最大。所以，乌越先生得出这样的结论：味道对综合满意度的影响最大。

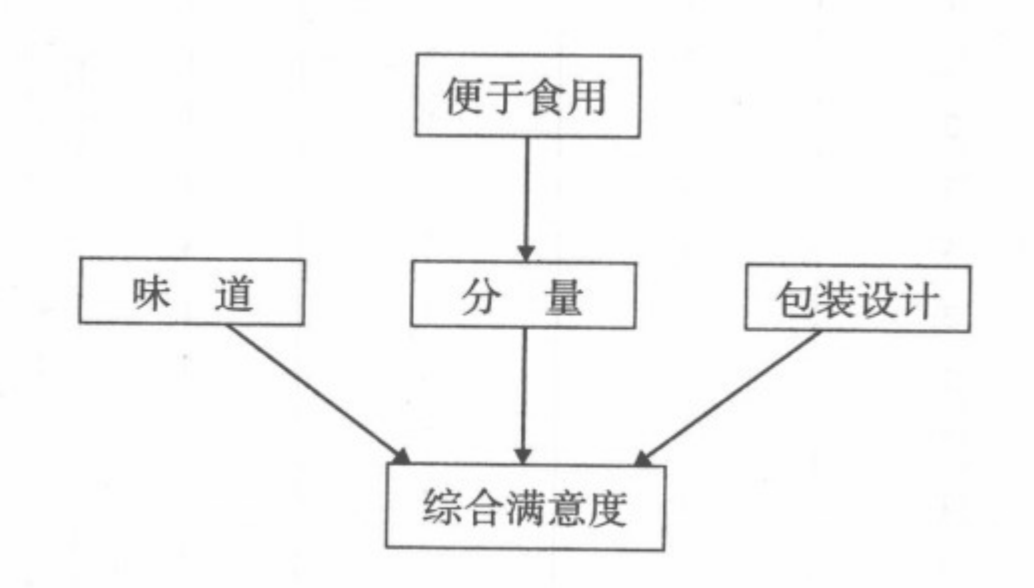
1. 在调查“各自变量对因变量的影响”时使用的一种方法。
2. 变量标准化以后推导出的重回归方程的偏回归系数，被称为标准偏回归系数。

味道对综合满意度的影响最大，鸟越先生得出了这样的结论。我们可以先不去管鸟越先生的心情如何，但是他的这种想法是值得我们关注的。

鸟越先生认定，上表中的各变量存在着如下关系：



这和我们认定的重回归分析结构[1]是一致的。但这也未必可行，也许真实的情况会是：



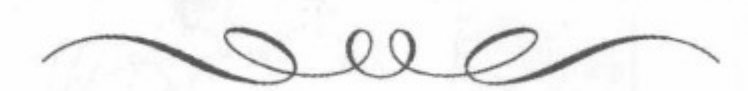
存在这样的关系也不是不可能的。

同重回归分析相比，研究“各自变量对因变量的影响”更倾向于结构方程式模型[2]这种分析方法。但是结构方程式模型，是一种“各自变量对因变量的影响”可以“自动地”明确判定出来的分析方法，而不是一种没有规则的分析方法。这种分析方法需要分析者在分析前，主观地假定各变量之间的关系，之后才能求出路径系数[3]。

1. 请参见 105 页。
2. 通常将其称为协方差结构分析。
3. 相当于重回归分析中的偏回归系数或标准偏回归系数。

◆第 4 章◆

Logistic 回归分析







1. Logistic回归分析
老爸让我买的咖啡豆。
还要买风见面包店的牛角面包……
然后……
啊！对了，
约翰的点心！
嗯……
KAZAMI
是他!
ドキ


ドキ
怎么办……
跟他说话……
ドキ
KAZAMI
ドキ
但是，
还是不敢！！

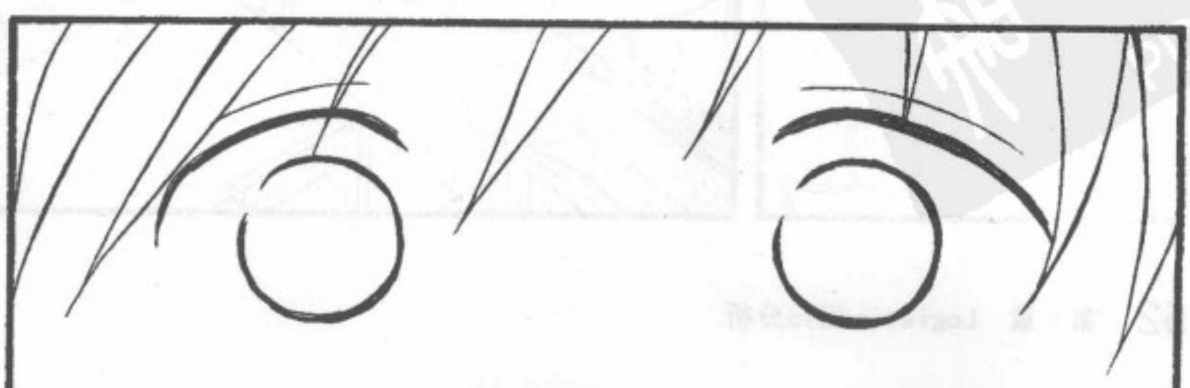

加油啊！


ドキ
ドキ
ぎゅっ
是啊！
如果不振作
的话……

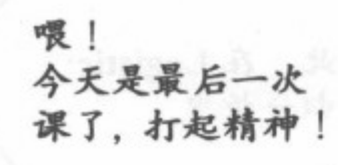

从现在起，要学会坚强，
所以没关系的……


ドキ
ドキ
对了，可以从他
落下的那本书聊
起……


ドキ
要锁定……




那……
那个……
TeaRoom NORNS
哎呦！
美羽！
啊！
是！
为什么唉声叹气的？今天不是要学 Logistic 回归分析？
为了美羽，才从开店前开始讲课的！
对，对不起……


喂！
今天是最后一次
课了，打起精神！


咦！


喂！
啊……


也许是，
认错人了
吧……
谢谢你！

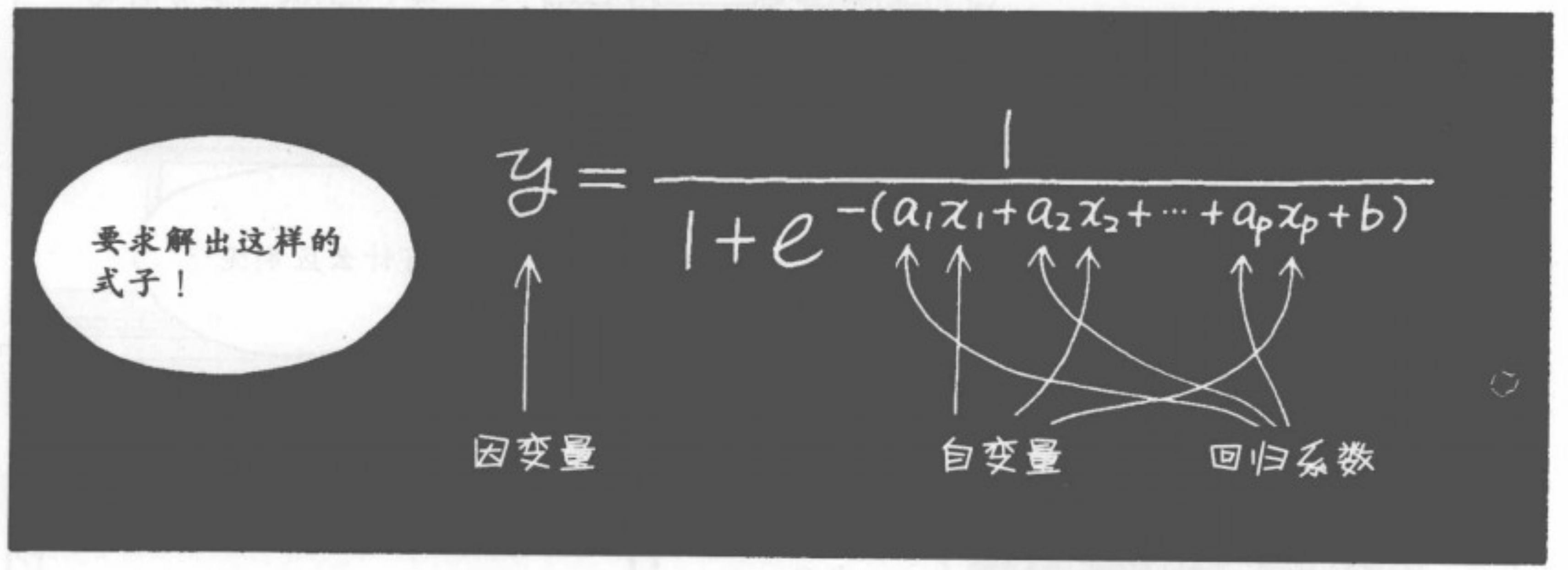

无论是回归分析，还是重回归分析，都是预测"冰红茶销售量"或者"营业额"这样的数值的分析方法。
而Logistic回归分析就不一样了。
有什么区别呢？

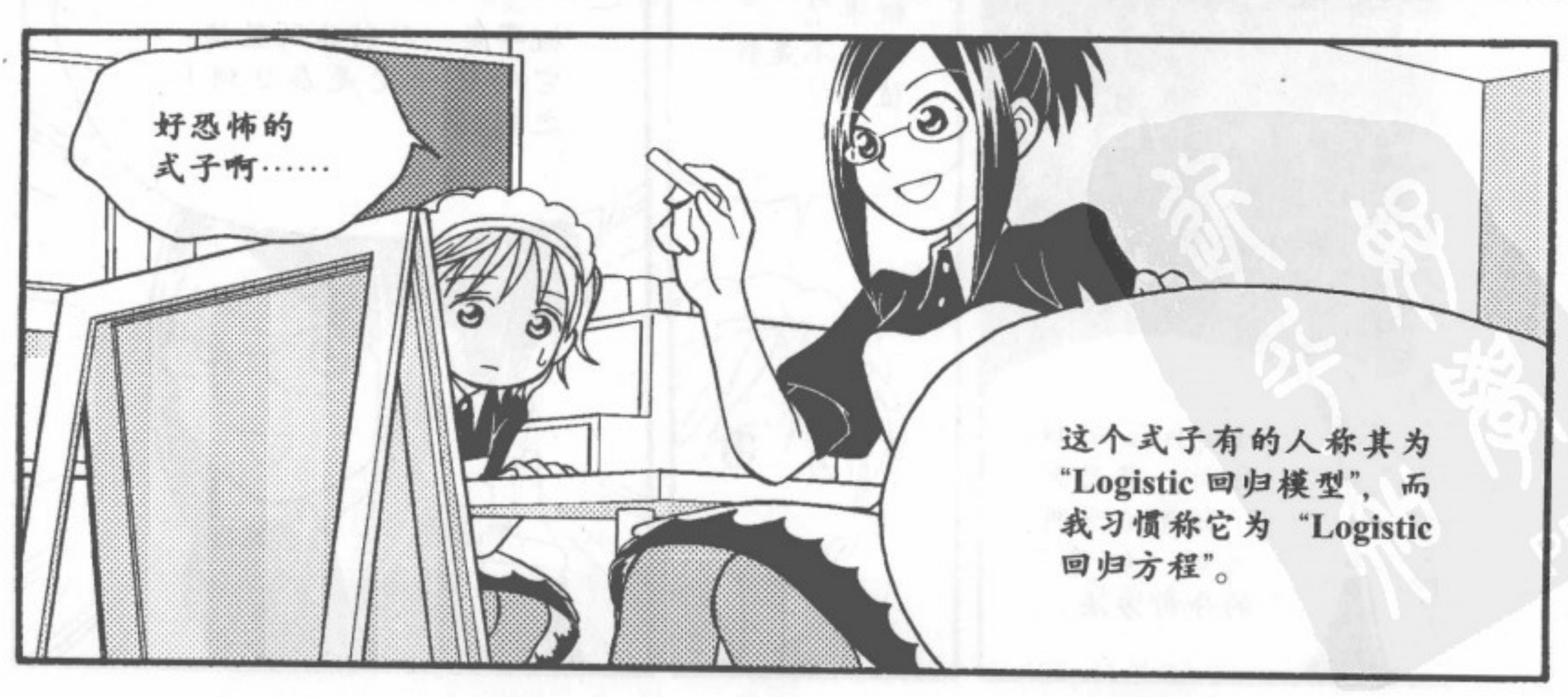

·○○君考进××大学的概率
·△△君患癌症的概率
Logistic回归分析，就是为了预测这里所说的"概率"的分析方法。

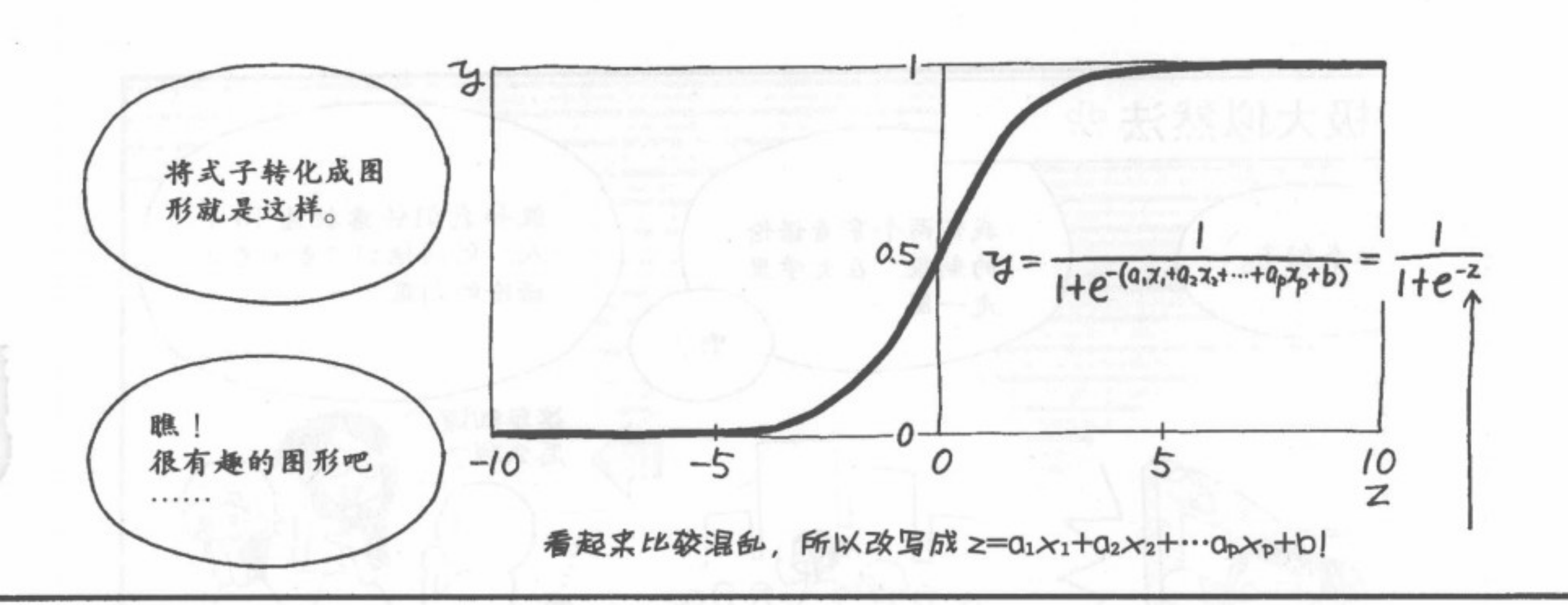

概率？
难道不是数值吗？
?


概率是一种特殊的数值，它的值一定是在0到1之间。


啊、是这样啊……
回归分析和重回归分析是没有办法预测介于0到1之间的预测值的。
所以只能用 Logistic 回归分析。
因此，在 Logistic 回归分析中，
カッカッ

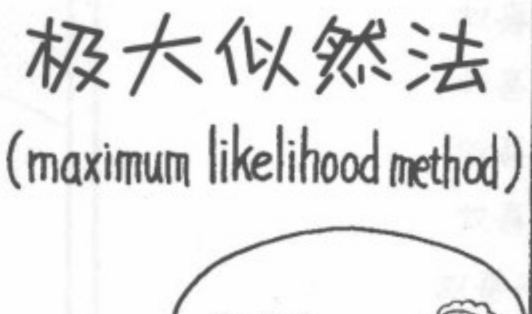

要求解出这样的式子！
$y = \frac{1}{1+e^{-(a_1x_1+a_2x_2+\cdots+a_px_p+b)}}$
因变量
自变量
回归系数


好恐怖的式子啊……
这个式子有的人称其为"Logistic 回归模型"，而我习惯称它为"Logistic 回归方程"。


将式子转化成图形就是这样。
瞧！很有趣的图形吧……
y
1
0.5
0
-10
-5
0
5
10
z
$y=\frac{1}{1+e^{-(a_1x_1+a_2x_2+\cdots+a_px_p+b)}}=\frac{1}{1+e^{-z}}$
看起来比较混乱，所以改写成 $z=a_1x_1+a_2x_2+\cdots a_px_p+b$！


喏！无论 z 的值是多少，y 的值总是位于 0 和 1 之间，对吧？
啊，真是这样啊！


接下来，为了便于理解 *Logistic* 回归分析，我们有必要先了解极大似然法的相关知识，
极大似然法？


它的主要思想是把“使样本出现的可能性最大”作为决策的准则。
极大似然法
(maximum likelihood method)


啊哈哈……


那，我们就先从这个话题开始吧！
拜托您了！

✿2. 极大似然法✿

举个例子。

我们两个穿着诺伦的制服，在大学里走一圈……

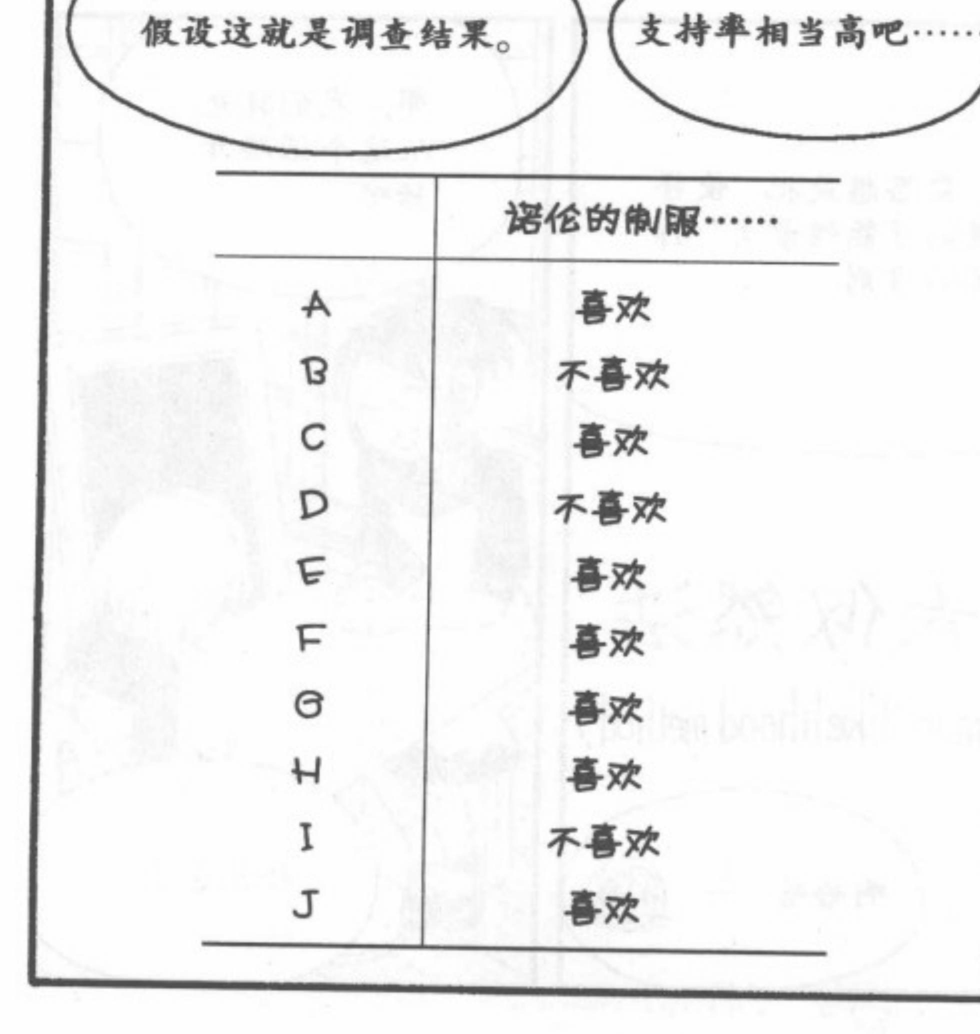
啊！

假如我们任意挑选 10 个人，问问他们“喜不喜欢诺伦的制服？”

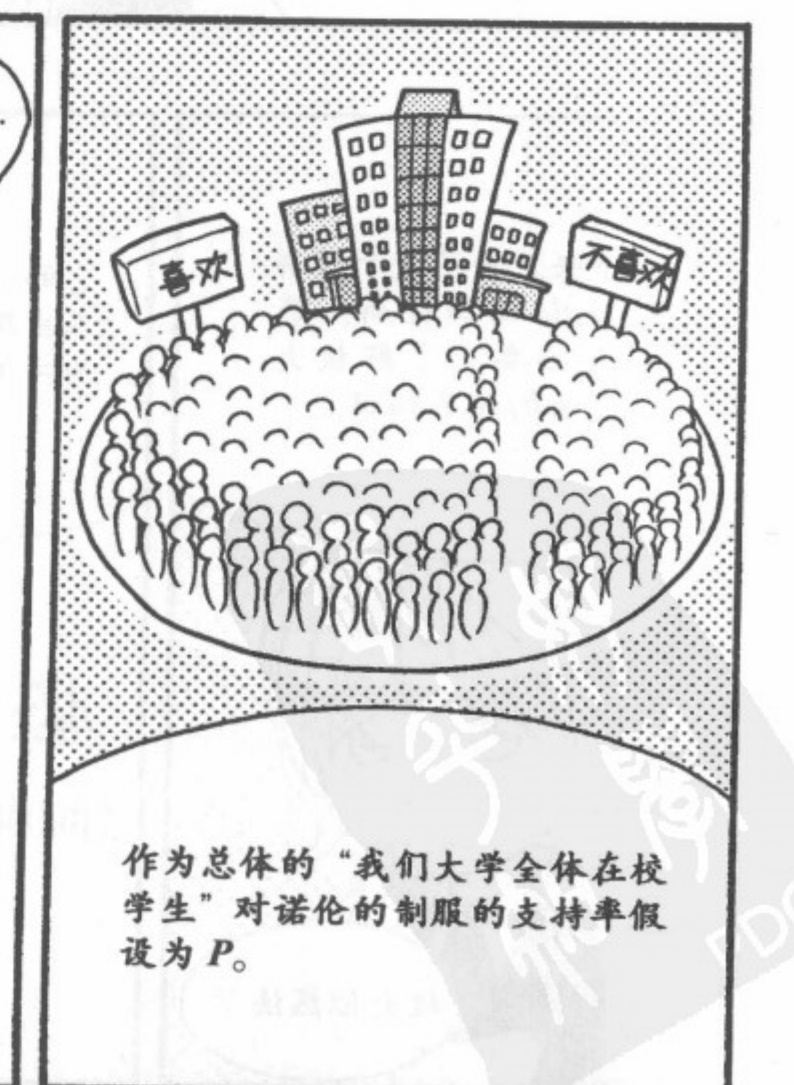
这身制服怎么样？

……嗯

有点难为情啊……



	诺伦的制服……
A	喜欢
B	不喜欢
C	喜欢
D	不喜欢
E	喜欢
F	喜欢
G	喜欢
H	喜欢
I	不喜欢
J	喜欢



像刚刚表中的那种情况，发生的概率应该是这样的。

喜欢 不喜欢 喜欢 不喜欢 喜欢 喜欢 喜欢 喜欢 不喜欢 喜欢

$$p \times (1-p) \times p \times (1-p) \times p \times p \times p \times p \times (1-p) \times p$$

$$= p^7(1-p)^3$$

嗯……



作为总体的“我们大学全体在校学生”对诺伦的制服的支持率 p 的值，一定会令

$p^7(1-p)^3$

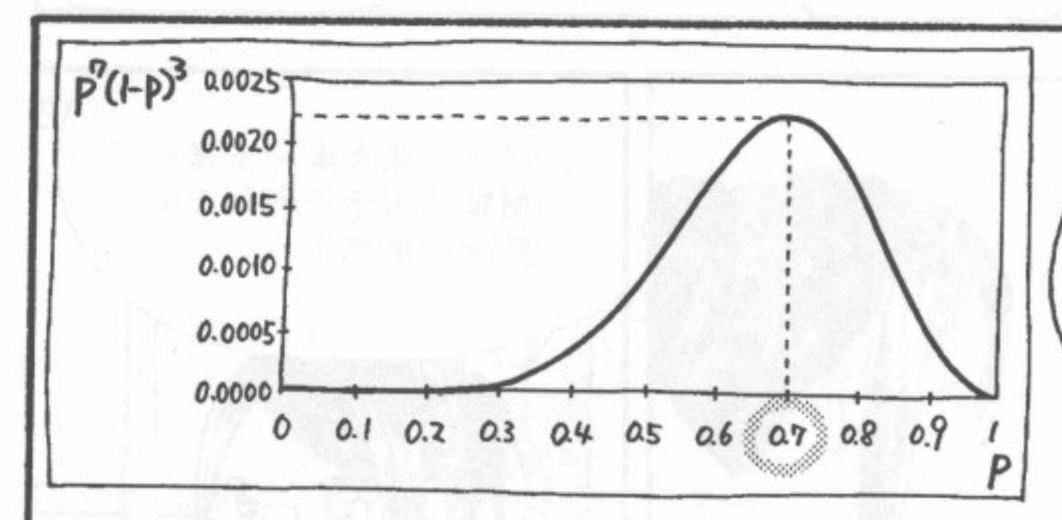
或者

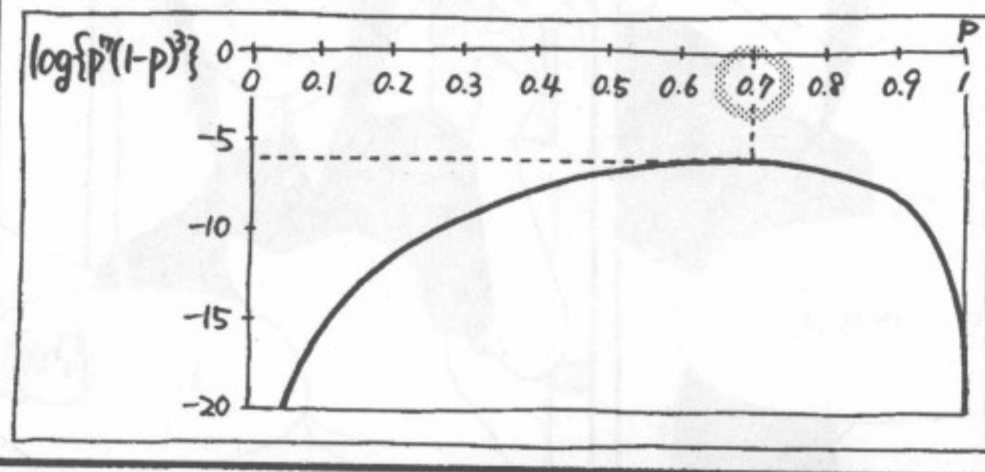
$\log\{p^7(1-p)^3\}$

的值为最大。

它的基本想法，可以这样解释。

嗯……

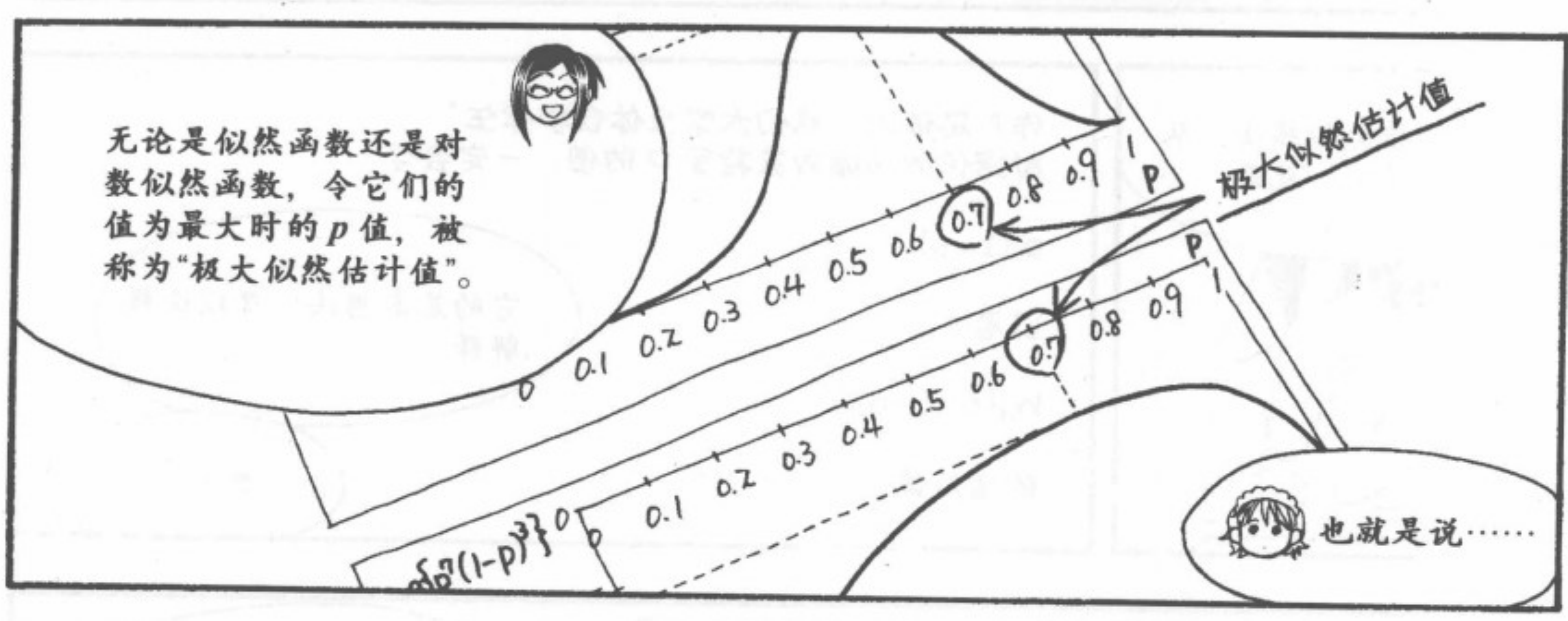




从图像上解释的话，就是求这些图像最高点所对应的横轴坐标的一种想法。




$p^7(1-p)^3$ 称为“似然函数”。
$\log\{p^7(1-p)^3\}$ 称为“对数似然函数”。
$p^7(1-p)^3$
→ 似然函数
$\log\{p^7(1-p)^3\}$
→ 对数似然函数


无论是似然函数还是对数似然函数，令它们的值为最大时的 p 值，被称为“极大似然估计值”。
极大似然估计值
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1 p
也就是说……


极大似然法是求解极大似然估计值的一种方法？
可以那样说！
那么，就来求一下诺伦制服的例子中的极大似然估计值吧！
好！

步骤 1　求似然函数。

$$p \times (1-p) \times p \times (1-p) \times p \times p \times p \times p \times (1-p) \times p$$
$$= p^7(1-p)^3$$

步骤 2　求对数似然函数，并整理。

$$\begin{aligned} L &= \log\{p^7(1-p)^3\} \\ &= \log p^7 + \log(1-p)^3 \\ &= 7\log p + 3\log(1-p) \end{aligned}$$

此后，将对数似然函数记为 L。



步骤 3　对数似然函数 L 关于 p 求微分，令其值为 0。

$$\frac{dL}{dp} = 7 \times \frac{1}{p} + 3 \times \frac{1}{1-p} \times (-1) = 7 \times \frac{1}{p} - 3 \times \frac{1}{1-p} = 0$$

步骤 4　整理步骤 3 中的式子，求出极大似然估计值。

$$7 \times \frac{1}{p} - 3 \times \frac{1}{1-p} = 0$$

$$\left(7 \times \frac{1}{p} - 3 \times \frac{1}{1-p}\right) \times p(1-p) = 0 \times p(1-p)$$ ← 两边同时乘以 $p(1-p)$

$$7(1-p) - 3p = 0$$

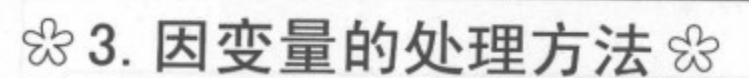
$$7 - 7p - 3p = 0$$

$$7 - 10p = 0$$

$$p = \frac{7}{10}$$







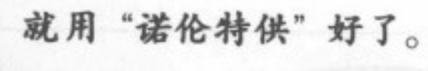
3. 因变量的处理方法



Tea Room NORNS
那么，我们再回到Logistic回归分析。



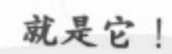
我们用什么数据来分析呢？
嗯，嗯，嗯……



就用“诺伦特供”好了。







啊……



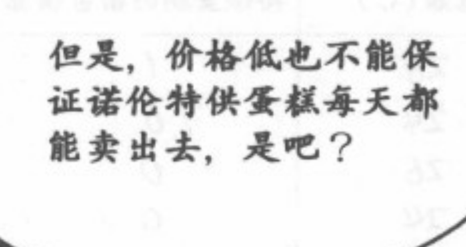
Tea Room NORNS

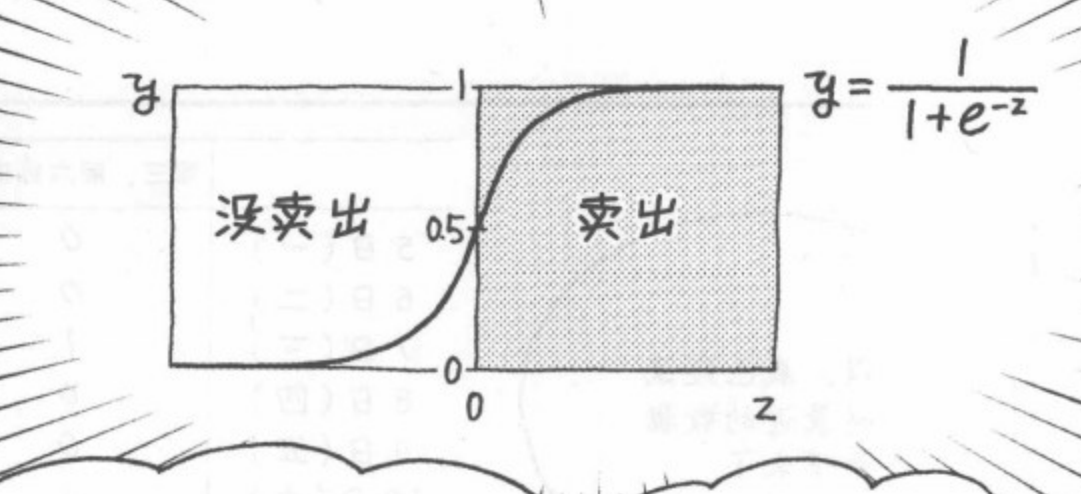
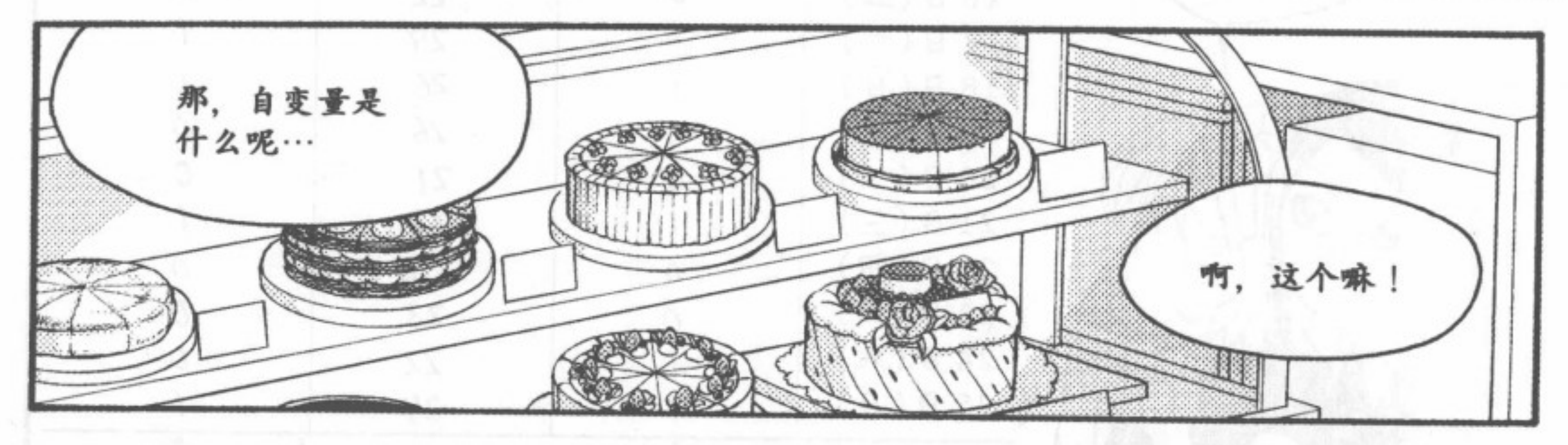
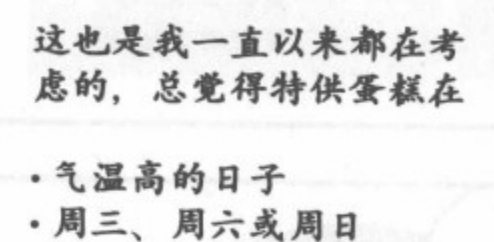


诺伦特供，就是店里的诺伦特供蛋糕吗？
诺伦特供，每天仅限1个，7000日元。



就是它！
基本看老板的心情才会卖出的蛋糕。



但是，价格低也不能保证诺伦特供蛋糕每天都能卖出去，是吧？
说来也是…

y
1
0.5
0
0
z
没卖出
卖出
$y = \frac{1}{1+e^{-z}}$
既然如此，我们就来求一下预测卖出概率的Logistic回归方程吧！
哈哈！这对我们店也是很有用的呢！

那，自变量是什么呢…
啊，这个嘛！


这也是我一直以来都在考虑的，总觉得特供蛋糕在
·气温高的日子
·周三、周六或周日
比较好卖。
会是那样吗？

嗯，本来周六、日光顾的客人就比较多，
ワーッ
而周三是因为附近大学的社团成员经常来诺伦聚会，这样，人也显得比较多！

所以，我已经试着将最近的数据记录下来了。

好厉害哦！理纱前辈。

	周三、周六或周日	最高气温(℃)	特供蛋糕的销售情况
5日(一)	0	28	1
6日(二)	0	24	0
7日(三)	1	26	0
8日(四)	0	24	0
9日(五)	0	23	0
10日(六)	1	28	1
11日(日)	1	24	0
12日(一)	0	26	1
13日(二)	0	25	0
14日(三)	1	28	1
15日(四)	0	21	0
16日(五)	0	22	0
17日(六)	1	27	1
18日(日)	1	26	1
19日(一)	0	26	0
20日(二)	0	21	0
21日(三)	1	21	1
22日(四)	0	27	0
23日(五)	0	23	0
24日(六)	1	22	0
25日(日)	1	24	1

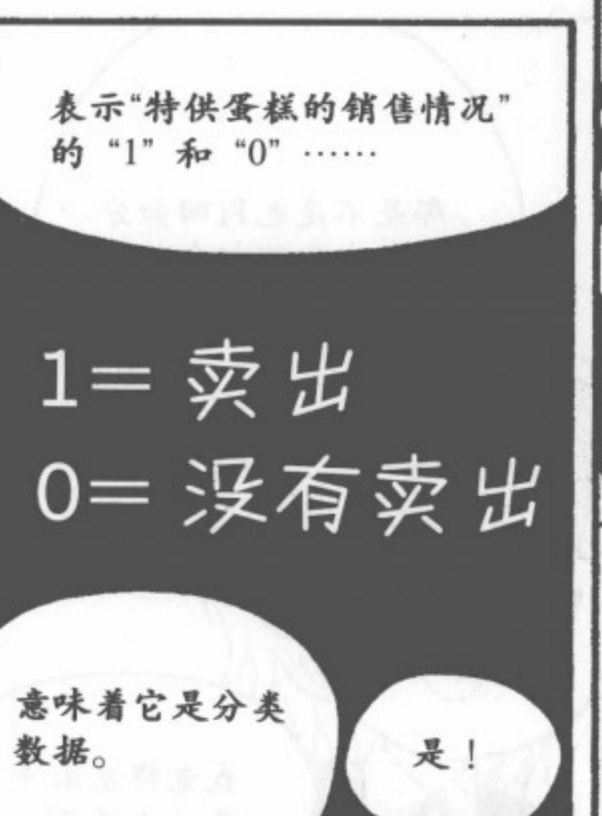
"1"是指周三、周六或周日
"0"是指其他日子

"1"是卖出去了
"0"是没有卖出去









啊，“周三、周六
或周日”也是分类
数据。

没错！

本来是分类数据的，
把它看成数值数据。

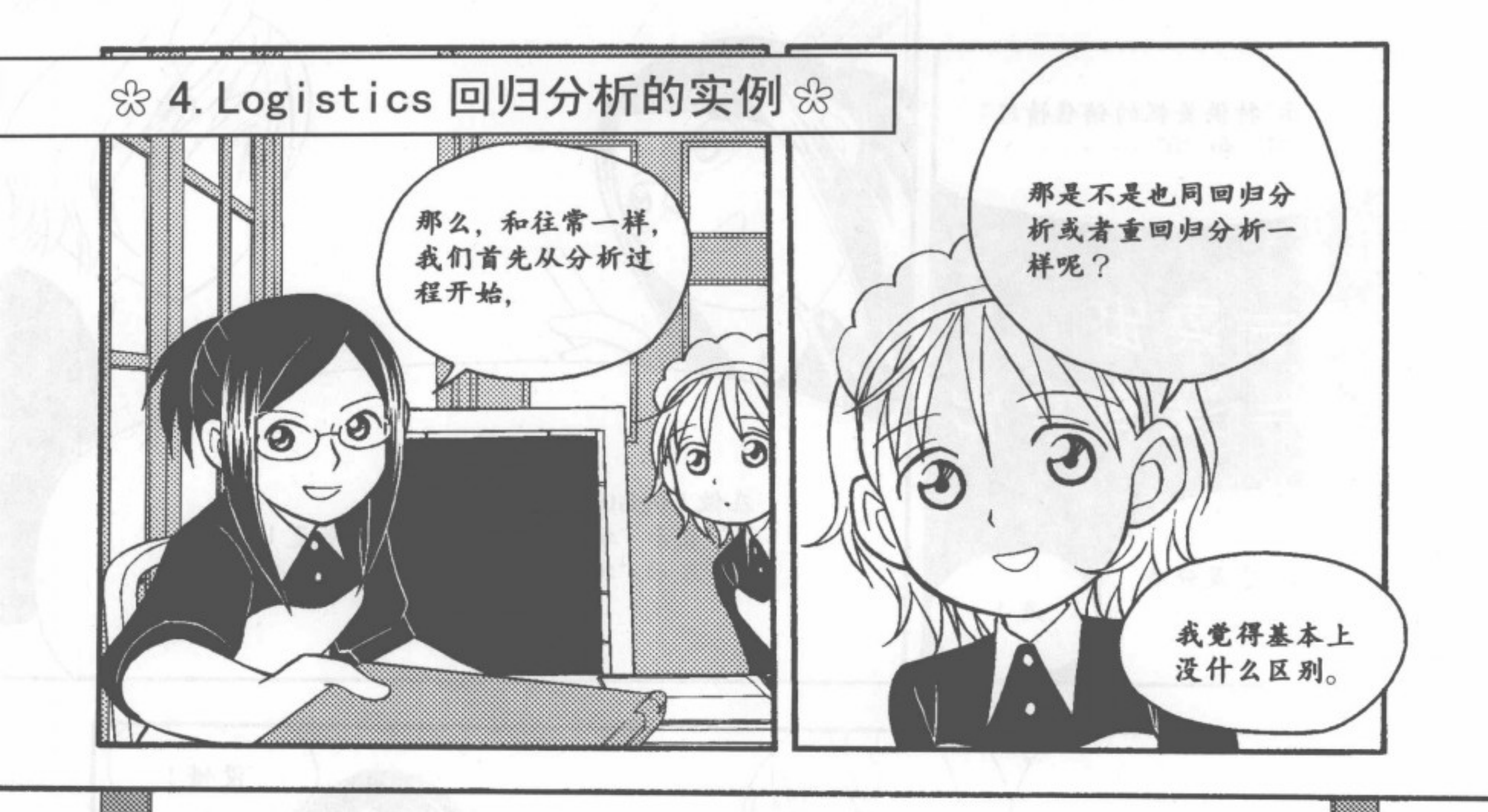
在 Logistic 回归分析中，
自变量可以是

- 仅为数值数据
- 仅为分类数据
- 数值数据和分类数据的混合

无论哪一种，都可以进行分析。

和重回归分析一样。

✿ 4. Logistics 回归分析的实例 ✿



Logistic 回归分析的过程：

① 首先，为了讨论是否具有求解 Logistic 回归方程的意义，画出各个自变量和因变量的散点图。

↓

② 求解 Logistic 回归方程。

↓

③ 确认 Logistic 回归方程的精度。

↓

④ 进行“回归系数的检验”。

↓

⑤ 预测。

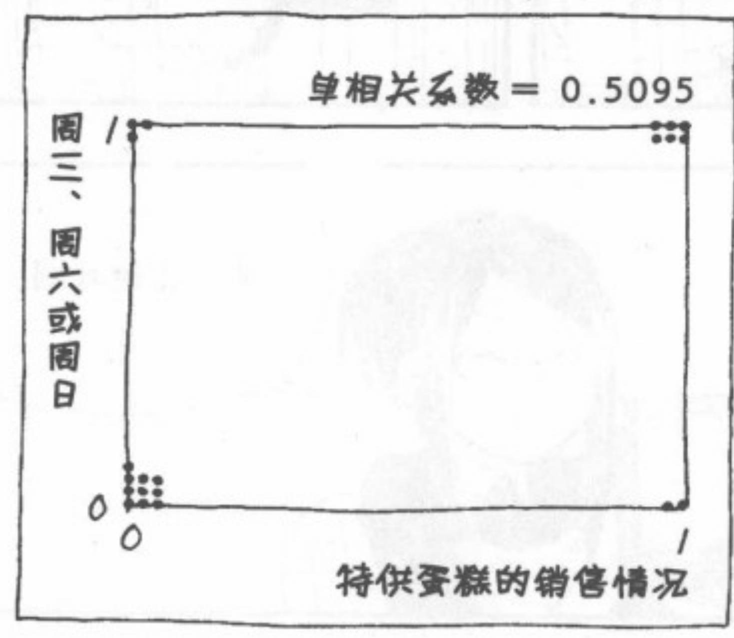
这就是 Logistic 回归分析的过程。

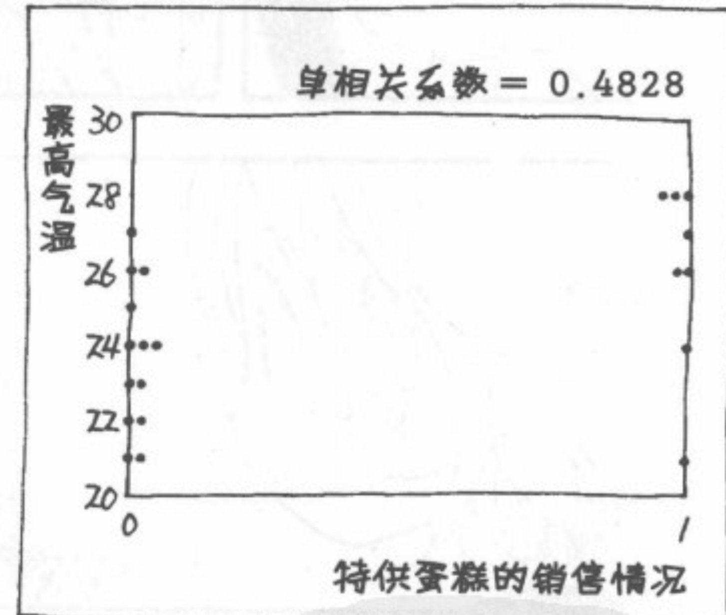
知道了。

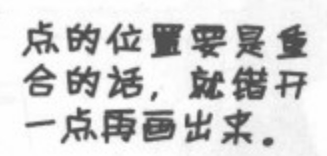
① 首先，为了讨论是否具有求解 Logistic 回归方程的意义，画出各个自变量和因变量的散点图。











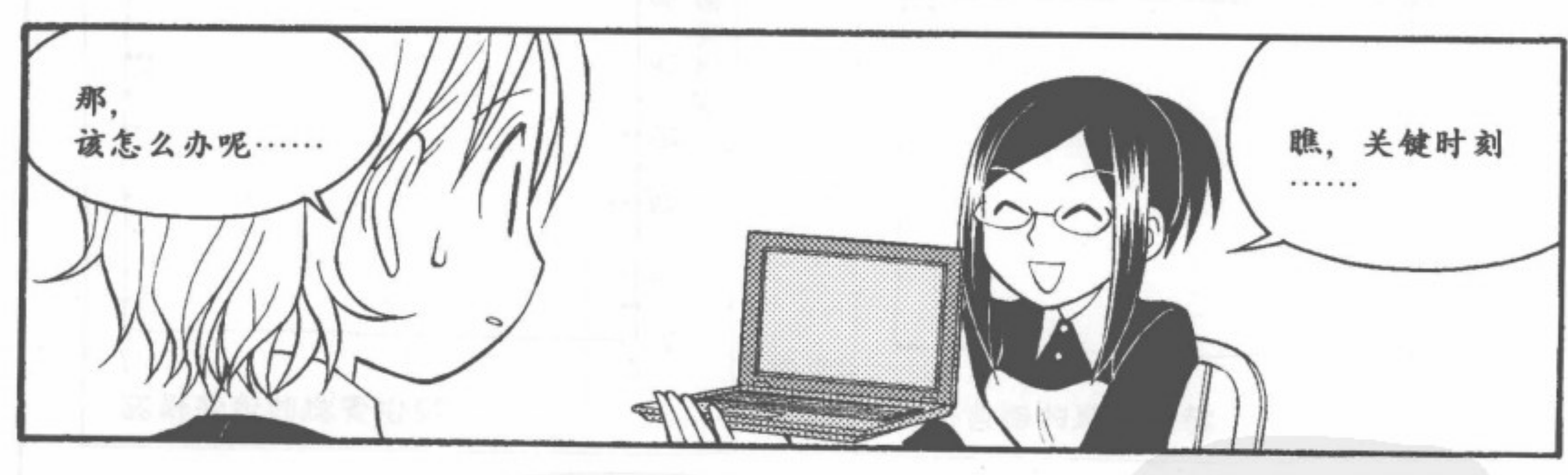


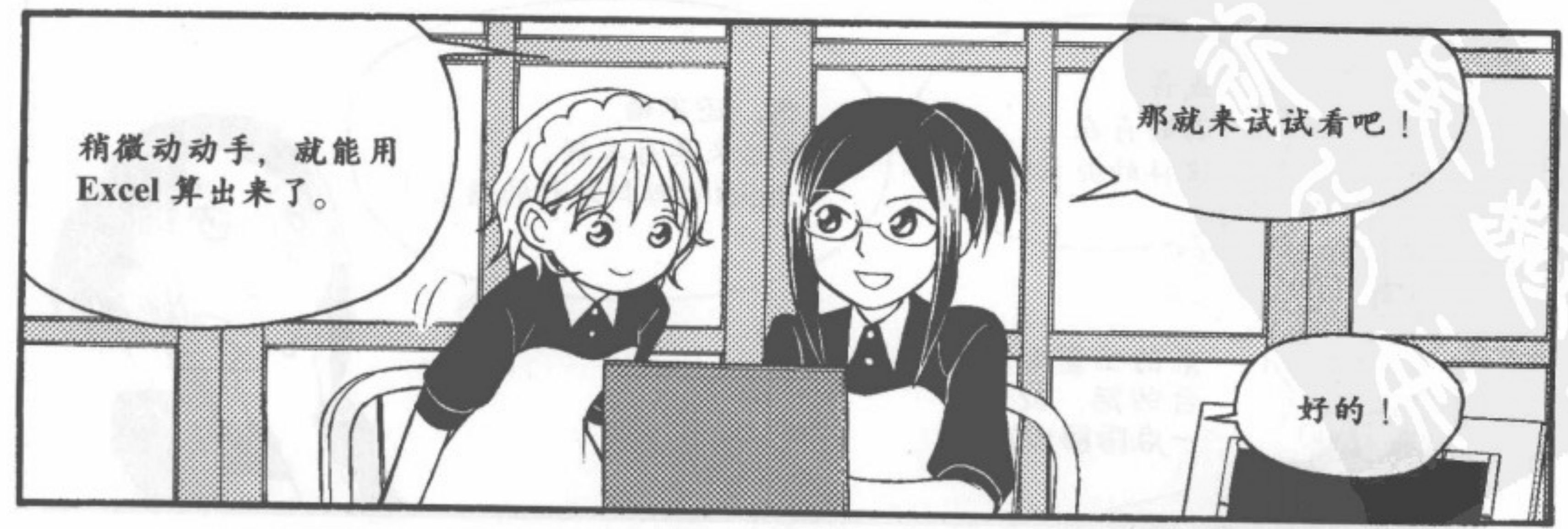
② 求解 Logistic 回归方程











步骤 1 计算过程如下表所述：

	周三、周六或周日 x_1	最高温度 x_2	特供蛋糕的销售情况 y	特供蛋糕的销售情况 $\hat{y}=\frac{1}{1+e^{-(a_1x_1+a_2x_2+b)}}$
5日(一)	0	28	1	$\frac{1}{1+e^{-(a_1\times 0+a_2\times 28+b)}}$
6日(二)	0	24	0	$\frac{1}{1+e^{-(a_1\times 0+a_2\times 24+b)}}$
⋮ ⋮	⋮	⋮	⋮	⋮
25日(日)	1	24	1	$\frac{1}{1+e^{-(a_1\times 1+a_2\times 24+b)}}$

步骤 2 步骤 2 求解似然函数。

$$\underbrace{\frac{1}{1+e^{-(a_1\times 0+a_2\times 28+b)}}}_{\text{卖出}}\times\underbrace{\left[1-\frac{1}{1+e^{-(a_1\times 0+a_2\times 24+b)}}\right]}_{\text{没有卖出}}\times\cdots\times\underbrace{\frac{1}{1+e^{-(a_1\times 1+a_2\times 24+b)}}}_{\text{卖出}}$$

步骤 3 步骤 3 求解对数似然函数。

$$L=\log\left\{\frac{1}{1+e^{-(a_1\times 0+a_2\times 28+b)}}\times\left[1-\frac{1}{1+e^{-(a_1\times 0+a_2\times 24+b)}}\right]\times\cdots\times\frac{1}{1+e^{-(a_1\times 1+a_2\times 24+b)}}\right\}$$

$$=\log\left[\frac{1}{1+e^{-(a_1\times 0+a_2\times 28+b)}}\right]+\log\left[1-\frac{1}{1+e^{-(a_1\times 0+a_2\times 24+b)}}\right]+\cdots+\log\left[\frac{1}{1+e^{-(a_1\times 1+a_2\times 24+b)}}\right]$$

步骤 4 求解极大似然估计值。

极大似然估计值，就是使对数似然函数 L 的值最大时 a_1、a_2 和 b 的值。

$$\begin{cases} a_1 = 2.44 \\ a_2 = 0.54 \\ b = -15.20 \end{cases}$$



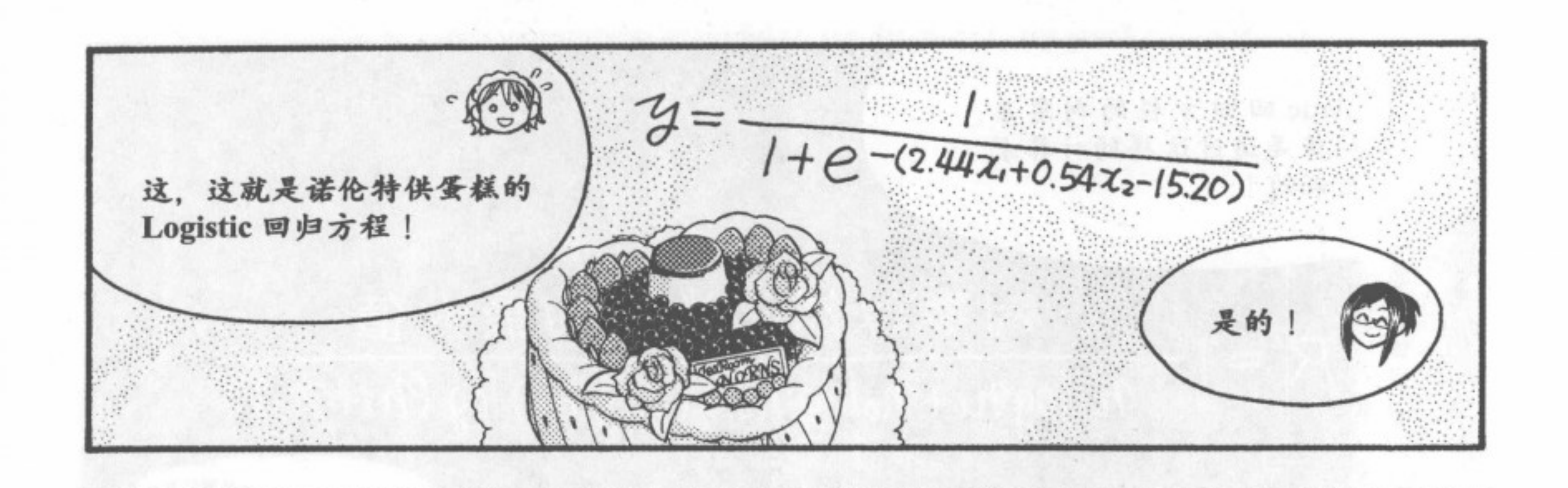
求解方法请参照第 208 页。
对数似然函数 L 的最大值，虽然和步骤 4 没有直接关系，但是考虑到其重要性，在这里先讲一下。如下所示：

$$L = \log\left[\frac{1}{1+e^{-(2.44\times 0+0.54\times 28-15.20)}}\right] + \log\left[1-\frac{1}{1+e^{-(2.44\times 0+0.54\times 24-15.20)}}\right] + \cdots + \log\left[\frac{1}{1+e^{-(2.44\times 1+0.54\times 24-15.20)}}\right]$$
$$= -8.9$$

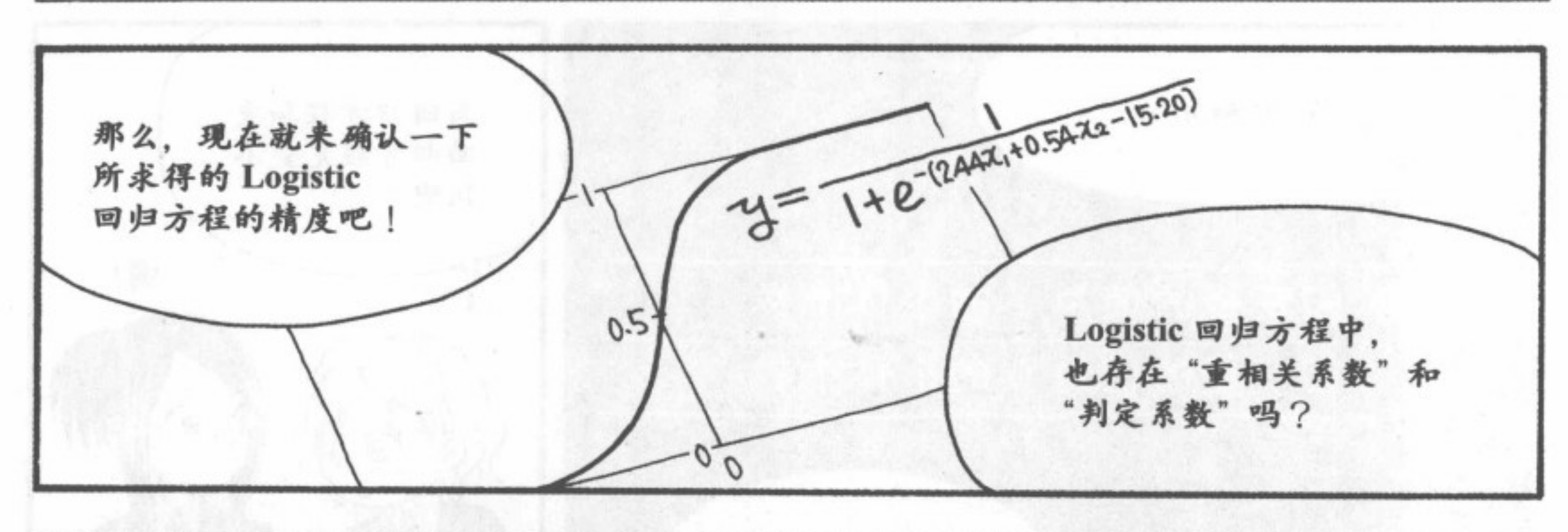
步骤 5 求解 Logistic 回归方程。

对比步骤 4 可知，Logistic 回归方程为

$$y = \frac{1}{1+e^{-(2.44x_1+0.54x_2-15.20)}}$$



③ 确认 Logistic 回归方程的精度







Logistic 回归方程的判定系数的值是通过这样的计算求解出来的！

$$R^2 = 1 - \frac{\text{对数似然函数 L 的最大值}}{n_1 \log n_1 + n_0 \log n_0 - (n_1 + n_0)\log(n_1 + n_0)}$$



哇……

式子中的 n_1 和 n_0……

n_1	因变量的值为 1 的个体个数
n_0	因变量的值为 0 的个体个数

是这样的意思。



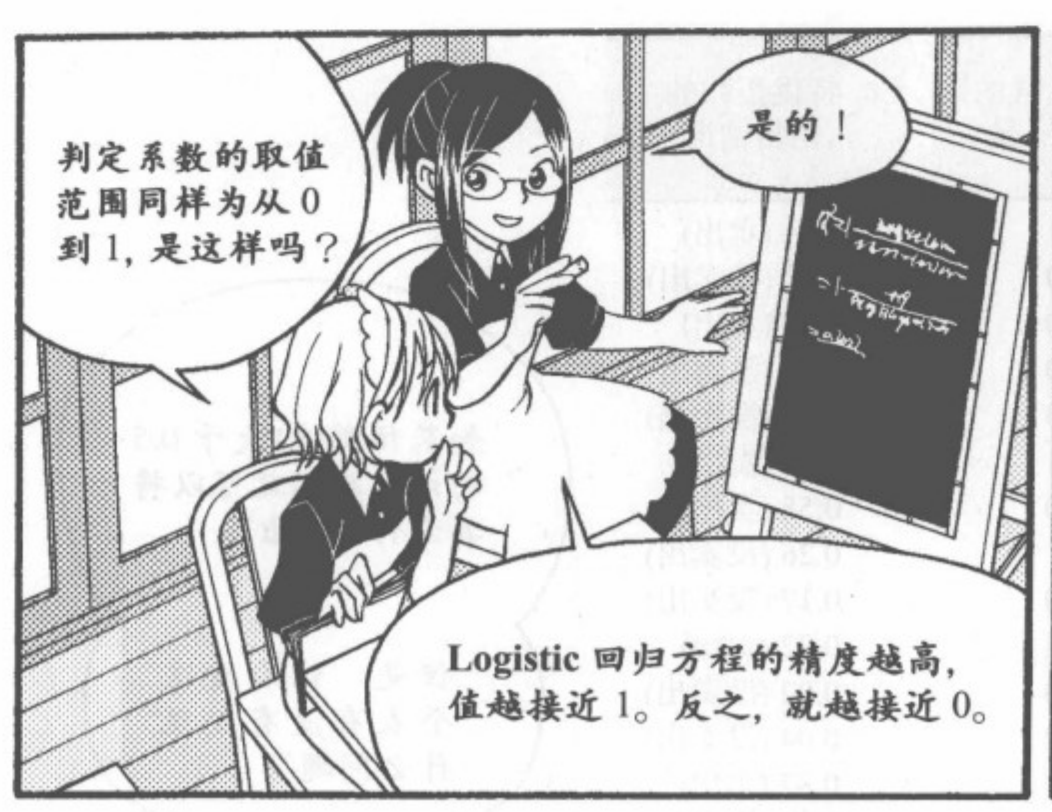
那么，你就来具体计算一下吧！

$$R^2 = 1 - \frac{\text{对数似然函数 L 的最大值}}{n_1 \log n_1 + n_0 \log n_0 - (n_1 + n_0)\log(n_1 + n_0)}$$

$$= 1 - \frac{-8.9}{8\log 8 + 13\log 13 - (8+13)\log(8+13)}$$

$$= 0.3622$$

唉，想不到会这么低。


判定系数的取值范围同样为从0到1，是这样吗？
是的！
Logistic 回归方程的精度越高，值越接近1。反之，就越接近0。


也同回归方程一样，没有标准吗？
没错！


尽管如此，Logistic 回归方程的判定系数的值基本上都不是很大，所以，参考程度十分有限。
是吗……


我再教你其他的求解精度的方法。
カチャ
カチャ
其他的方法？


看这个表。
？

	周三、周六或周日 x_1	最高气温 x_2	特供蛋糕的销售情况 y	特供蛋糕的销售情况 $\hat{y}$
5日(一)	0	28	1	0.51 (卖出)
6日(二)	0	24	0	0.11 (没卖出)
7日(三)	1	26	0	0.80 (卖出)
8日(四)	0	24	0	0.11 (没卖出)
9日(五)	0	23	0	0.06 (没卖出)
10日(六)	1	28	1	0.92 (卖出)
11日(日)	1	24	0	0.58 (卖出)
12日(一)	0	26	1	0.26 (没卖出)
13日(二)	0	25	0	0.17 (没卖出)
14日(三)	1	28	1	0.92 (卖出)
15日(四)	0	21	0	0.02 (没卖出)
16日(五)	0	22	0	0.04 (没卖出)
17日(六)	1	27	1	0.87 (卖出)
18日(日)	1	26	1	0.80 (卖出)
19日(一)	0	26	0	0.26 (没卖出)
20日(二)	0	21	0	0.02 (没卖出)
21日(三)	1	21	1	0.21 (没卖出)
22日(四)	0	27	0	0.38 (没卖出)
23日(五)	0	23	0	0.06 (没卖出)
24日(六)	1	22	0	0.31 (没卖出)
25日(日)	1	24	1	0.58 (卖出)

$$\frac{1}{1+e^{-(2.44\times1+0.54\times24-15.20)}}=0.58$$



啊！7日和11日明明是没有卖出的，但是$\hat{y}$却显示卖出了。



	y	$\hat{y}$
7日(三)	0	0.80（卖出）
11日(日)	0	0.58（卖出）

没错，还有呢？



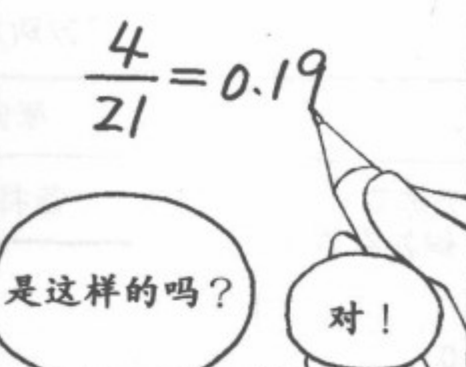
12日和21日明明是卖出了，但是$\hat{y}$却显示没有卖出。





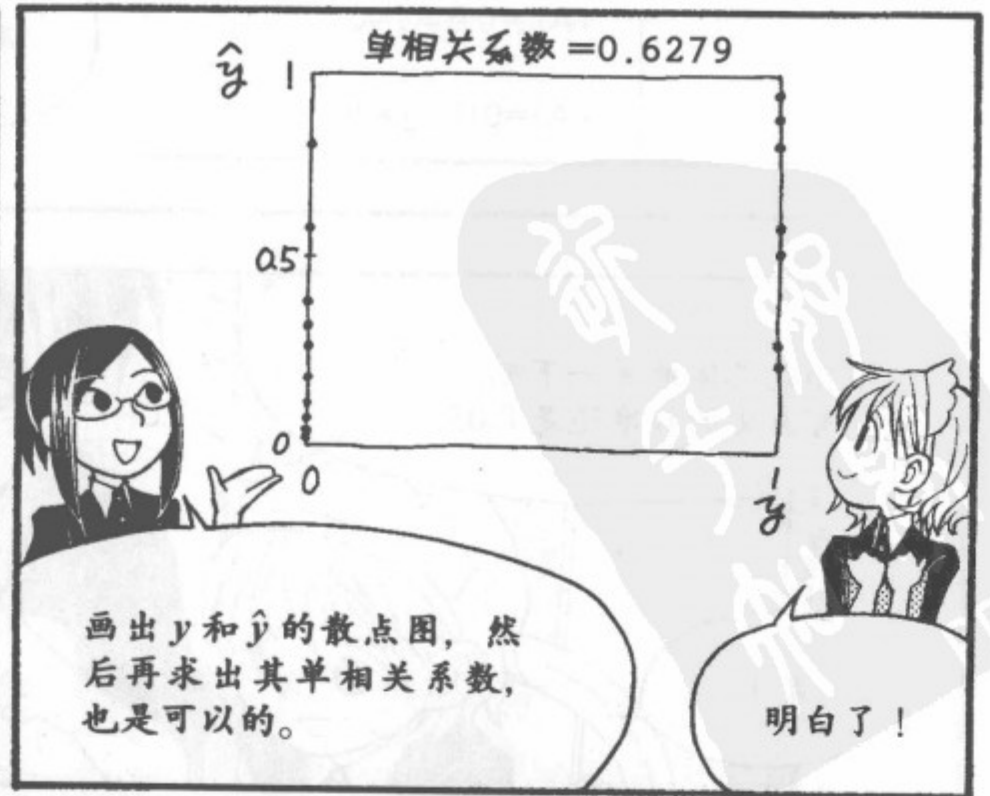

用来确认 Logistic 回归方程
精度的“误判率”是……
分析结果与事实不一致的个体个数
全部个体个数
这样定义的！


也可以叫做
“未中率”吧？
嗯！


那么，这个例子中的
误判率……
4/21=0.19
是这样的吗？
对！


这个值越小，Logistic
回归方程的精度越高！


啊，对了，还有另外一个可
以确认 Logistic 回归方程精
度的方法……


单相关系数=0.6279
ŷ
1
0.5
0
0
1
y
画出 y 和 ŷ 的散点图，然
后再求出其单相关系数，
也是可以的。
明白了！

④ 进行“回归系数的检验”





“全面讨论偏回归系数的检验”

原假设	“$A_1 = A_2 = 0$”。
备择假设	“$A_1 = A_2 = 0$”不成立 即，下面任意一组关系成立 · $A_1 \neq 0$ 且 $A_2 \neq 0$ · $A_1 \neq 0$ 且 $A_2 = 0$ · $A_1 = 0$ 且 $A_2 \neq 0$

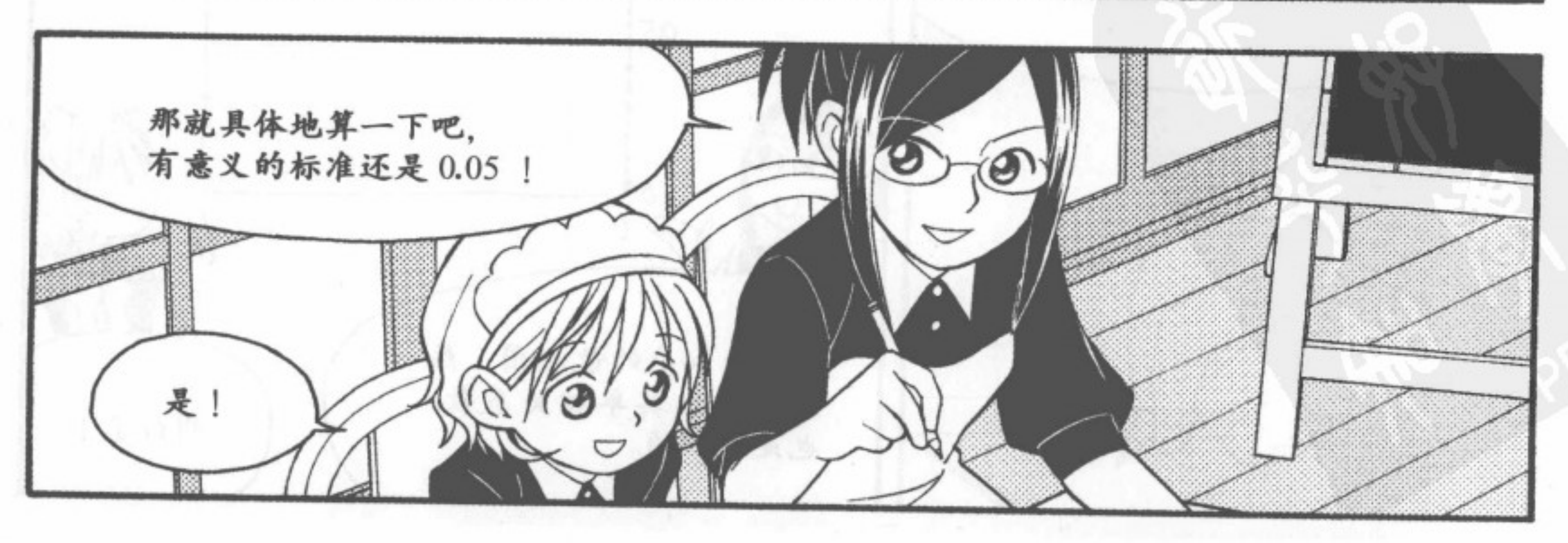
“分别讨论偏回归系数的检验”

原假设	“$A_i = 0$”。
备择假设	“$A_i \neq 0$”

是的！

重回归分析时的笔记本

是这样吗？





首先，进行“全面讨论偏回归系数的检验”。顺便提一下，通过下述计算进行的检验，通常被称为“似然比检验”（Likelihood Ratio Test）。

步骤 1	定义总体。	将“周三、周六或周日为 x_1 、最高气温为 x_2℃的日子”作为总体。
步骤 2	建立原假设和备择假设。	原假设为“ $A_1=A_2=0$ 成立”。 备择假设为“ $A_1=A_2=0$ 不成立”。
步骤 3	选择所要进行的“检验”类型。	进行“全面讨论偏回归系数的检验”。
步骤 4	设定有意义的标准。	以 0.05 为有意义的标准。
步骤 5	通过样本数据求出检验统计量的值。	下面进行“全面讨论偏回归系数的检验”。 “全面讨论偏回归系数的检验”的检验统计量为 2[对数似然函数的最大值 $-n_1\log n_1-n_0\log n_0+(n_1+n_0)\log(n_1+n_0)$] 所以，在本例题中的检验统计量的值为 $2[-8.9010-8\log 8-13\log 13+(8+13)\log(8+13)]=10.1$ 在本例题中，如果原假设成立，那么检验统计量就服从自由度为 2(= 自变量的个数)的 χ^2 分布※。
步骤 6	再将步骤 5 中求出的检验统计量的值所对应的 P 值，与有意义的标准进行比较，看看 P 值是否比其小。	有意义的标准是 0.05。检验统计量的值为 10.1，所以 P 值为 0.006。 $0.006<0.05$。所以 P 值较小。
步骤 7	如果在步骤 6 中，P 值比有意义的标准小，我们就可以得出“备择假设成立”的结论。反之，我们就可以得出“原假设成立”的结论。	与有意义的标准相比，P 值较小。所以，备择假设成立。

※ χ^2 分布中 P 值的求解方法请参见第 201 页。

接下来，进行“分别讨论偏回归系数的检验”。
我们以 A_1 为检验对象，示范一下！
顺便提一下，通过下述计算进行的检验，通常被称为“Wald 检验”。



步骤 1	定义总体。	将“周三、周六或周日为 x_1、最高气温为 x_2 ℃的日子”作为总体。
步骤 2	建立原假设和备择假设。	原假设为“$A_1=0$ 成立”。 备择假设为“$A_1 \neq 0$ 成立”。
步骤 3	选择所要进行的“检验”类型。	进行“分别讨论偏回归系数的检验”。
步骤 4	设定有意义的标准。	以 0.05 为有意义的标准。
步骤 5	通过样本数据求出检验统计量的值。	下面进行“分别讨论偏回归系数的检验”。 “分别讨论偏回归系数的检验”的检验统计量为 $\frac{a_1^2}{S^{11}}$※。 所以在本例题中的检验统计量的值为 $\frac{2.44^2}{1.5388}=3.9$。 在本例题中，如果原假设成立，那么检验统计量就服从自由度为 1 的 χ^2 分布。
步骤 6	再将步骤 5 中求出的检验统计量的值所对应的 P 值，与有意义的标准进行比较，看看 P 值是否比其小。	有意义的标准是 0.05。检验统计量的值为 3.9。所以 P 值为 0.0489。0.0489<0.05，所以 P 值较小。
步骤 7	如果在步骤 6 中 P 值比有意义的标准小，则我们就可以得出“备择假设成立”的结论。反之，我们就可以得出“原假设成立”的结论。	与有意义的标准相比，P 值较小。 所以，备择假设“$A_1 \neq 0$”成立。

※ S^{11} 的求解方法见下页说明。

在有些参考资料中，不是依据 χ^2 分布，而是依据标准正态分布来讲解“回归系数的检验”。
这个问题从数学的角度解释起来比较困难，所以我们不做详细介绍。但是，无论依据哪种分布，其最终的结论都是相同的。



在步骤 5 中出现的 S^{11} 是这样求出来的。

周三、周六或周日　　最高气温

$$\left[\begin{pmatrix}0 & 0 & \cdots & 1\\ 28 & 24 & \cdots & 24\\ 1 & 1 & \cdots & 1\end{pmatrix}\begin{pmatrix}(5日的\hat{y})\times(5日的1-\hat{y}) & 0 & \cdots & 0\\ 0 & (6日的\hat{y})\times(6日的1-\hat{y}) & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & (25日的\hat{y})\times(25日的1-\hat{y})\end{pmatrix}\begin{pmatrix}0 & 28 & 1\\ 0 & 24 & 1\\ \vdots & \vdots & \vdots\\ 1 & 24 & 1\end{pmatrix}\right]^{-1}$$

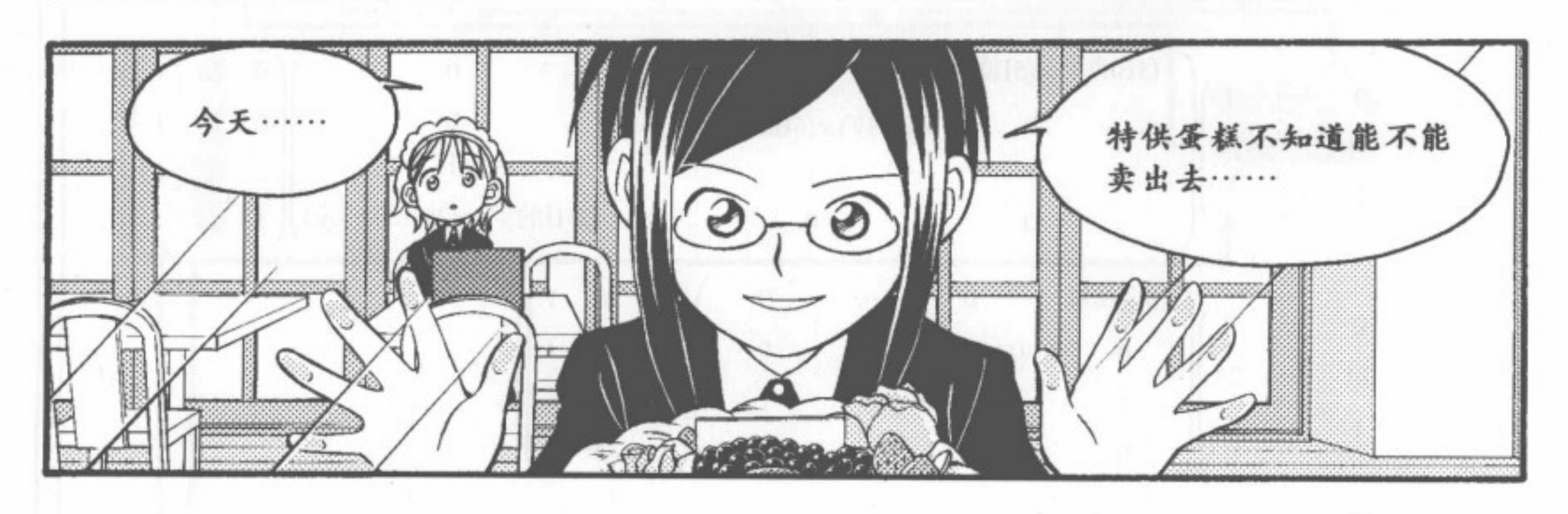
$$=\left[\begin{pmatrix}0 & 0 & \cdots & 1\\ 28 & 24 & \cdots & 24\\ 1 & 1 & \cdots & 1\end{pmatrix}\begin{pmatrix}0.51\times0.49 & 0 & \cdots & 0\\ 0 & 0.11\times0.89 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0.58\times0.42\end{pmatrix}\begin{pmatrix}0 & 28 & 1\\ 0 & 24 & 1\\ \vdots & \vdots & \vdots\\ 1 & 24 & 1\end{pmatrix}\right]^{-1}$$

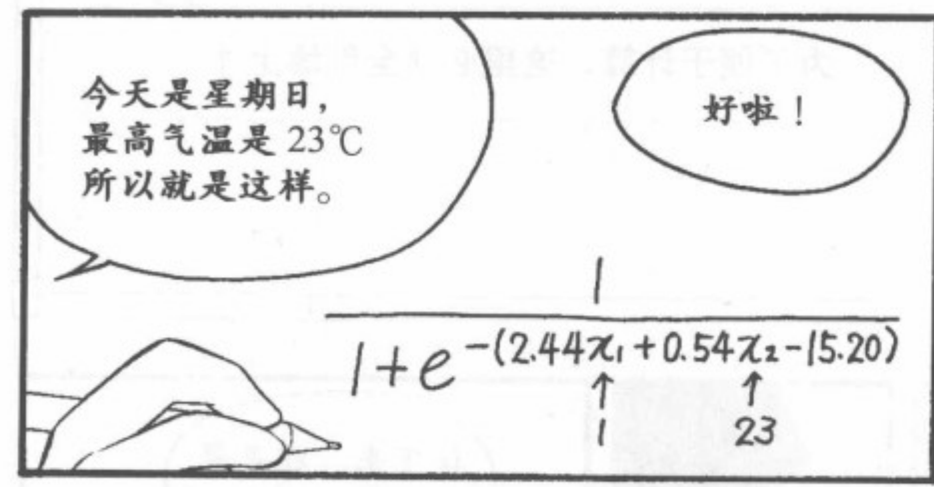
$$=\begin{pmatrix}1.5388 & \cdots & \cdots\\ \cdots & 0.0881 & \cdots\\ \cdots & \cdots & \cdots\end{pmatrix}$$

为了便于计算，这里必须全部填上 1。

这个就是 S^{11}。　这个就是 S^{22}。

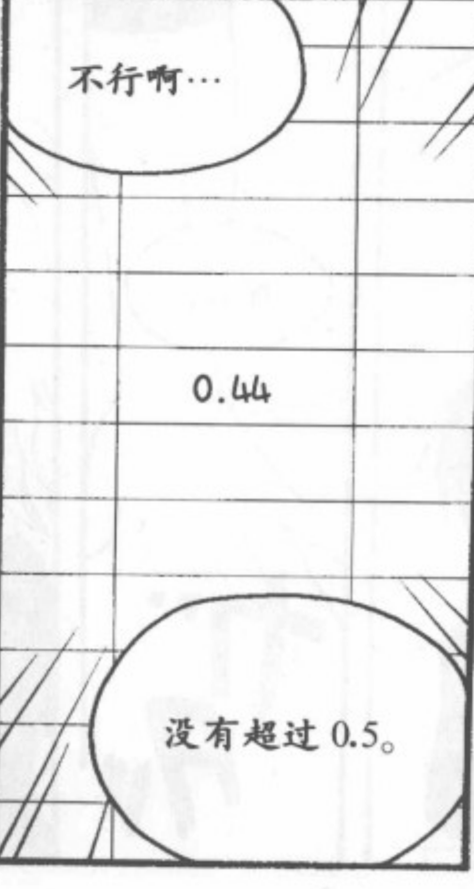
⑤ 预　测


今天……
特供蛋糕不知道能不能卖出去……


今天是星期日，最高气温是 23℃ 所以就是这样。
好啦！
$\frac{1}{1+e^{-(2.44x_1+0.54x_2-15.20)}}$
1
23


那么，就用计算机算一下看看吧！
是
カチャ
カチャ


啊…


不行啊…
0.44
没有超过 0.5。


今天卖不出去了！
可惜！
看起来只能这样了……



Tea Room NORNS
至此，有关“回归分析”中的
•回归分析
•重回归分析
•Logistic 回归分析
的学习，就告一段落了！
辛苦啦！
真的十分感谢！
还要继续努力啊！
这样，无论他什么时候来，你们不就能有共同话题了吗？
ぎゅっ

理、
理纱前辈……

你怎么对人那么亲切呢？

如果理纱前辈是我的对手的话，那我……

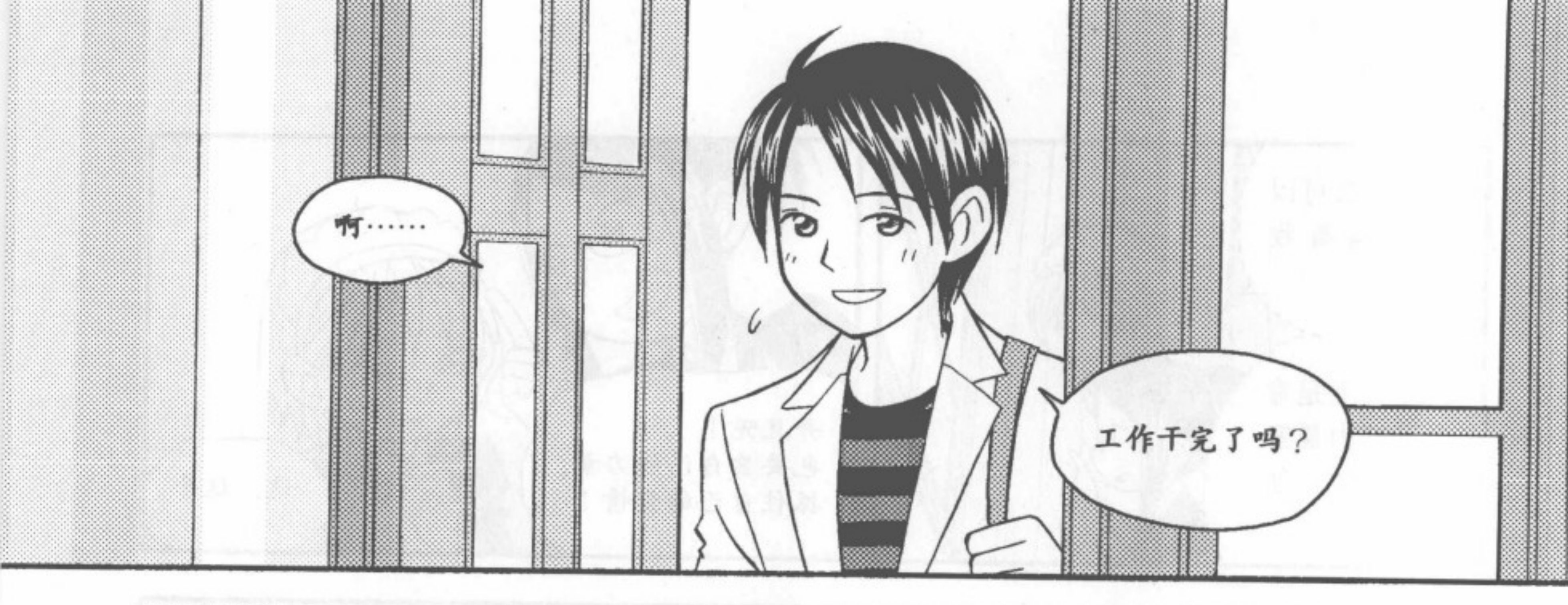
啊？

估计就没戏了。








啊……
工作干完了吗？

浩人！

浩人……
あわわ…

这是……
我的表弟浩人，
就住在邻街。

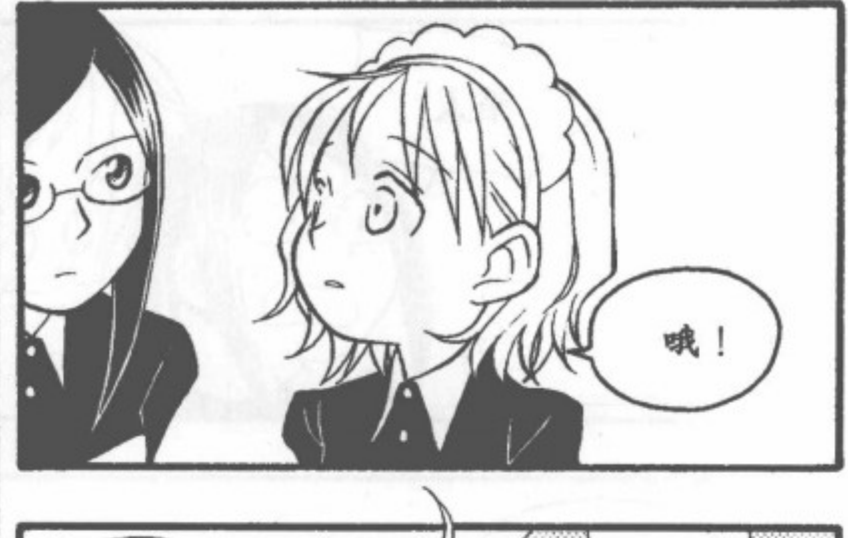
表弟？

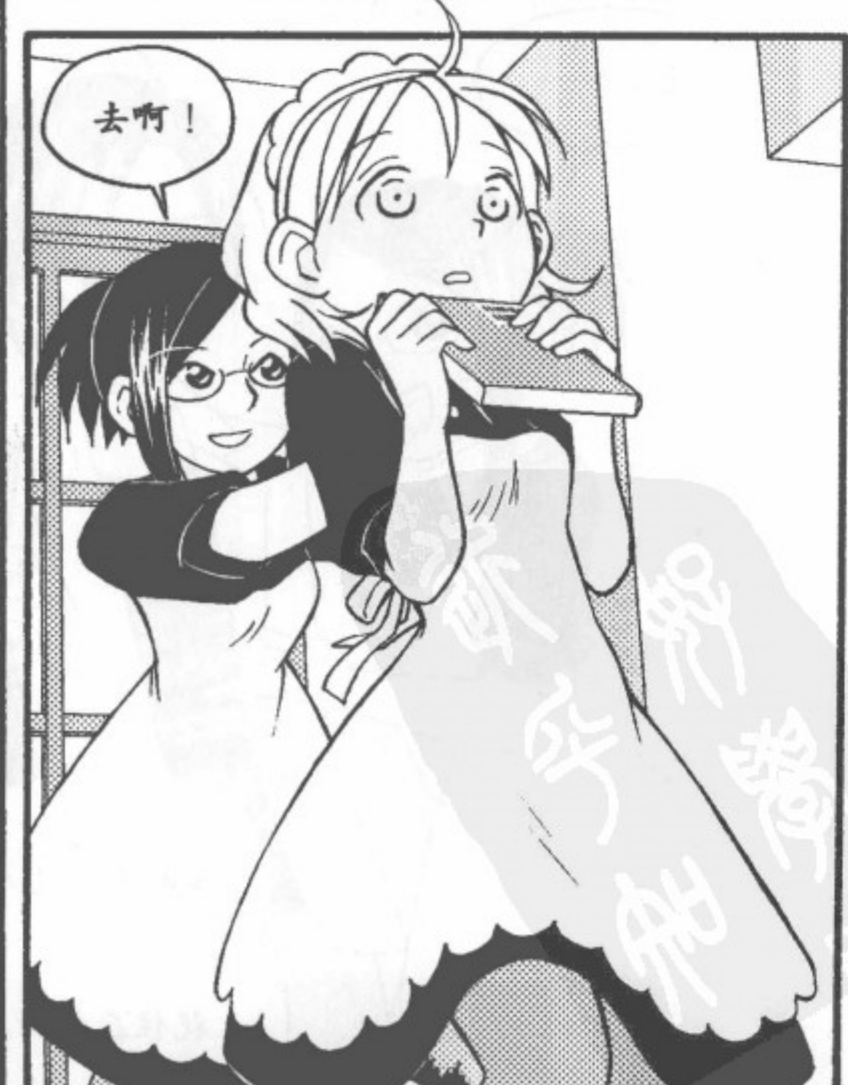
理，理纱前辈！
怎，怎么了……


怎，怎么可以一直瞒着我呢？
……美羽是靠自己的力量坚持努力的，所以……

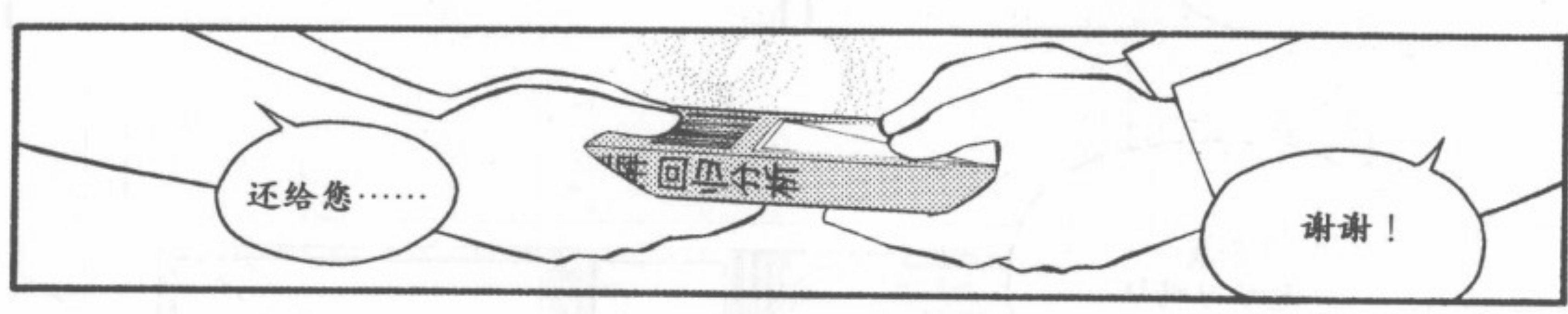

并且呢！也要靠自己的力量，抓住自己的爱情！
这，这个……


瞧！
人家可是特意从邻街跑过来见你的，正好可以把书还给他！


哦！


去啊！


嗯……
那个……


还给您……
谢谢！


是你一直在为我保管啊？
听理纱姐说。
さっ
嗯？


那，那个……
嗯……


嗯，我叫五十岚美羽……
我叫深泽浩人……
可以和我一起学习吗？
嗯！
谢谢！请多指教。


哈！
本店
马上开始
营业喽！




欢迎光临！


欢迎光临诺伦茶餐厅！

✿5. "Logistic回归分析过程"中的注意事项✿

下图再次给出第164页出现的"Logistic 回归分析过程"。

① 首先，为了讨论是否具有求解 Logistic 回归方程的意义，画出各个自变量和因变量的散点图。
↓
② 求解 Logistic 回归方程。
↓
③ 确认 Logistic 回归方程的精度。
↓
④ 进行"回归系数的检验"。
↓
⑤ 预测。

图 4.1 Logistic 回归分析过程

此前，在我们的讲解中，必须完成上图中的第①步到第④步。但事实并非如此，同回归分析、重回归分析一样，不同的情况下，只完成第①步到第③步就可以了。

✿6. Odds Ratio（优势比）✿

本节的内容较为抽象。如果是在本书中初次接触 Logistic 回归分析的读者，可以跳过这一部分，不做阅读。但是，从事与医疗相关领域的读者，还请您稍作了解。

6.1 Odds 和 Logit

$y=\frac{1}{1+e^{-z}}$ 可以改写成为 $\frac{y}{1-y}=e^{z}$ 或 $\log\frac{y}{1-y}=z$



$$y=\frac{1}{1+e^{-z}}=\frac{1}{1+e^{-z}}\times\frac{e^z}{e^z}=\frac{e^z}{e^z+1}$$

$$y\times(e^z+1)=\frac{e^z}{e^z+1}\times(e^z+1)$$ 两边同乘以(e^z+1)

$$y\times e^z+y=e^z$$

$$y=e^z-y\times e^z$$ 移项

$$y=(1-y)e^z$$

$$y\times\frac{1}{1-y}=(1-y)e^z\times\frac{1}{1-y}$$ 两边同乘以$\frac{1}{1-y}$

$$\frac{y}{1-y}=e^z$$

$$\log\frac{y}{1-y}=\log e^z=z$$

于是，第 168 页所得到的 $y=\frac{1}{1+e^{-(2.44x_1+0.54x_2-15.20)}}$ 可以改写为：

$$\frac{y}{1-y}=e^{2.44x_1+0.54x_2-15.20}$$

$$\log\frac{y}{1-y}=2.44x_1+0.54x_2-15.20$$

$\frac{y}{1-y}$叫做 Odds。$\log\frac{y}{1-y}$ 叫做 Logit。在本章的例子中，Odds 就是$=e^{2.44x_1+0.54x_2-15.20}$，

而 Logit 就是$2.44x_1+0.54x_2-15.20$。

6.2 优势比（Odds Ratio）和风险比（Risk Ratio）

下表再次给出第 162 页的表格。

◆表4.1 第162页的表格

	周三、周六或周日	最高气温(℃)	特供蛋糕的销售情况
5日(一)	0	28	1
6日(二)	0	24	0
7日(三)	1	26	0
8日(四)	0	24	0
9日(五)	0	23	0
10日(六)	1	28	1
11日(日)	1	24	0
12日(一)	0	26	1
13日(二)	0	25	0
14日(三)	1	28	1
15日(四)	0	21	0
16日(五)	0	22	0
17日(六)	1	27	1
18日(日)	1	26	1
19日(一)	0	26	0
20日(二)	0	21	0
21日(三)	1	21	1
22日(四)	0	27	0
23日(五)	0	23	0
24日(六)	1	22	0
25日(日)	1	24	1

下表是“周三、周六或周日”与“特供蛋糕的销售情况”的联列表（Cross-Tabulation Table）

◆表4.2 “周三、周六或周日”与“特供蛋糕的销售情况”的联列表

		特供蛋糕的销售情况		共计
		卖出	没卖出	
周三、周六或周日	是	6	3	9
	不是	2	10	12
共计		8	13	21

由这个联列表可知，“周三、周六或周日的卖出率”为$\frac{6}{9}$，“周三、周六或周日以外的卖出率”为$\frac{2}{12}$。

在实际操作中，往往会用到风险比的概念。

所谓风险比，以表 4.2 来说，就是：

$$\frac{\text{周三、周六或周日的卖出率}}{\text{周三、周六或周日以外的卖出率}}=\frac{\left(\frac{6}{9}\right)}{\left(\frac{2}{12}\right)}=\frac{6}{9}\div\frac{2}{12}=\frac{6}{9}\times\frac{12}{2}=\frac{2}{3}\times 6=4$$

在实际操作中，往往还会用到优势比的概念。

所谓优势比，以表 4.2 来说，就是：

$$\frac{\left(\frac{\text{周三、周六或周日的卖出率}}{1-\text{周三、周六或周日的卖出率}}\right)}{\left(\frac{\text{周三、周六或周日以外的卖出率}}{1-\text{周三、周六或周日以外的卖出率}}\right)}=\frac{\left[\frac{\left(\frac{6}{9}\right)}{1-\left(\frac{6}{9}\right)}\right]}{\left[\frac{\left(\frac{2}{12}\right)}{1-\left(\frac{2}{12}\right)}\right]}=\frac{\left[\frac{\left(\frac{6}{9}\right)}{\left(\frac{3}{9}\right)}\right]}{\left[\frac{\left(\frac{2}{12}\right)}{\left(\frac{10}{12}\right)}\right]}=\frac{\left[\frac{6}{3}\right]}{\left[\frac{2}{10}\right]}=\frac{6}{3}\div\frac{2}{10}=\frac{6}{3}\times\frac{10}{2}=2\times 5=10$$

以上的例子中，表面上看来，优势比虽然与风险比有些差别，但实际上，它们的值表示含义是相同的，所以，人们常常用优势比来替代风险比。

6.3 未调整的优势比和调整后的优势比

下表记录的是采用最优子集法对表 4.1 中的数据进行的分析。

◆表4.3 表4.1中数据的最优子集法分析结果

	自变量		Logistic回归方程	Odds
①	仅为“周三、周六或周日”	→	$y=\frac{1}{1+e^{-(2.30x_1-1.61)}}$	$e^{2.30x_1-1.61}$
②	仅为“最高气温”	→	$y=\frac{1}{1+e^{-(0.52x_2-13.44)}}$	$e^{0.52x_2-13.44}$
③	“周三、周六或周日”和“最高气温”	→	$y=\frac{1}{1+e^{-(2.44x_1+0.54x_2-15.20)}}$	$e^{2.44x_1+0.54x_2-15.20}$

“e 的‘情况①时的回归系数’次方”为 $e^{2.30}$，于是有

$$\frac{\text{周三、周六或周日的odds}}{\text{周三、周六或周日以外的odds}}=\frac{e^{2.30\times1-1.61}}{e^{2.30\times0-1.61}}=e^{2.30\times1-1.61-(2.30\times0-1.61)}=e^{2.30}$$

我们将其称为没有对“周三、周六或周日”进行调整的优势比。顺便说一下，$e^{2.30}=10$，这个值同前一页求得结果是一致的。

“e 的‘情况②时的回归系数’次方”为 $e^{0.52}$，于是有

$$\frac{\text{最高气温为}(k+1)\text{℃的odds}}{\text{最高气温为}k\text{℃的odds}}=\frac{e^{0.52\times(k+1)-13.44}}{e^{0.52\times k-13.44}}=e^{0.52\times(k+1)-13.44-(0.52\times k-13.44)}=e^{0.52}$$

我们将其称为没有对“最高气温”进行调整的优势比。

“e 的‘情况③时的回归系数’次方”为 $e^{2.44}$，于是有

$$\frac{e^{2.44\times1+0.54\times k-15.20}}{e^{2.44\times0+0.54\times k-15.20}}=e^{2.44\times1+0.54\times k-15.20-(2.44\times0+0.54\times k-15.20)}=e^{2.44}$$

我们将其称为对“周三、周六或周日”进行调整的优势比。

“e 的‘情况③时的回归系数’次方”为 $e^{0.54}$，于是有

$$\frac{e^{2.44\times1+0.54\times(k+1)-15.20}}{e^{2.44\times1+0.54\times k-15.20}}=\frac{e^{2.44\times0+0.54\times(k+1)-15.20}}{e^{2.44\times0+0.54\times k-15.20}}=e^{0.54\times(k+1)-15.20-(0.54\times k-15.20)}=e^{0.54}$$

我们将其称为对“最高气温”进行调整的优势比。

6.4 总体优势比的检验

在介绍 Logistic 回归分析的文献中，有时会提及“总体优势比的检验”。“总体优势比的检验”和第 176 页所讲过的“分别讨论偏回归系数的检验”一样。但是原假设和备择假设却有所不同。同“分别讨论偏回归系数的检验”中的

原假设	$A_i=0$
备择假设	$A_i\neq0$

相比，“总体优势比的检验”中则是：

原假设	$e^{A_i}=e^0=1$
备择假设	$e^{A_i}\neq e^0=1$

6.5 总体优势比的估计

一般情况下，之前所讲的“总体优势比的检验”的结果，会同总体优势比的置信区间一起介绍，至少在与医疗相关领域是这样的。所以下面来介绍一下总体优势比的置信区间的求解方法。

例如，在本章的例子中，当置信度为 95% 时，“周三、周六或周日”的总体优势比的置信区间可以通过以下计算求得。当置信度为 99% 时，只需将下图中的“1.96”换成“2.58”即可。

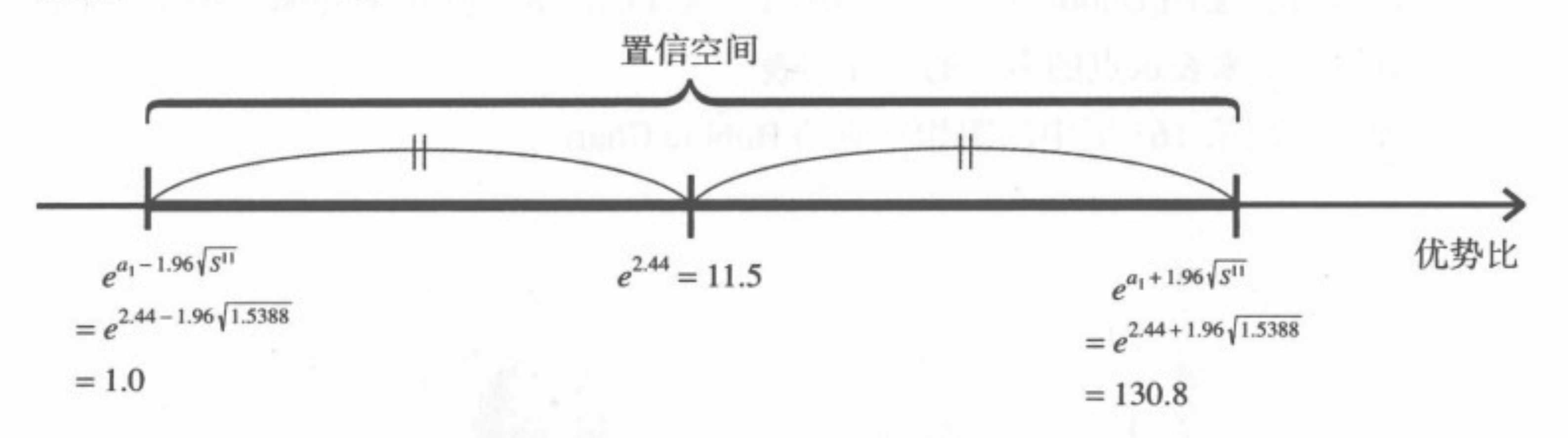


图4.2 置信度为95%时、“周三、周六或周日”的总体优势比的置信区间

※ a_1 的值请参照第 168 页。S^{11} 的值请参照第 177 页。

✿7. “检验”的名称 ✿

在“第 2 章 回归分析”、“第 3 章 重回归分析”和“第 4 章 Logistic 回归分析”中，出现了这么几种“检验”：回归系数的检验、偏回归系数的检验、全面讨论（偏）回归系数的检验、分别讨论（偏）回归系数的检验、总体优势比的检验。

这些名称是笔者自己的想法，并非通用的名称。这样做也是苦于找不到通用的名称。此外，似然比检验、wald 检验这些名称则是通用的名称，而非笔者自己的观点。

✿ 8. Bubble Chart（气泡图）✿

本节所讲的内容同 Logistic 回归分析没有任何关系，但是却比较有用，所以在这里稍作介绍。

在第 165 页中，美羽为了避免在同一个地方重复画点所使用的方法，是一个不错的主意。但是如果点的个数更多，例如 21 个甚至 210 个的话，就会弄得到处是点，最终变成一幅不知所谓的图，同不使用这种方法也没什么差别。

人们常常会使用 Bubble Chart（气泡图）这样的图表。所谓 Bubble Chart，就是通过气泡的大小来表示点的多少的一种图表。

下图就是同第 165 页中的图相对应的 Bubble Chart。

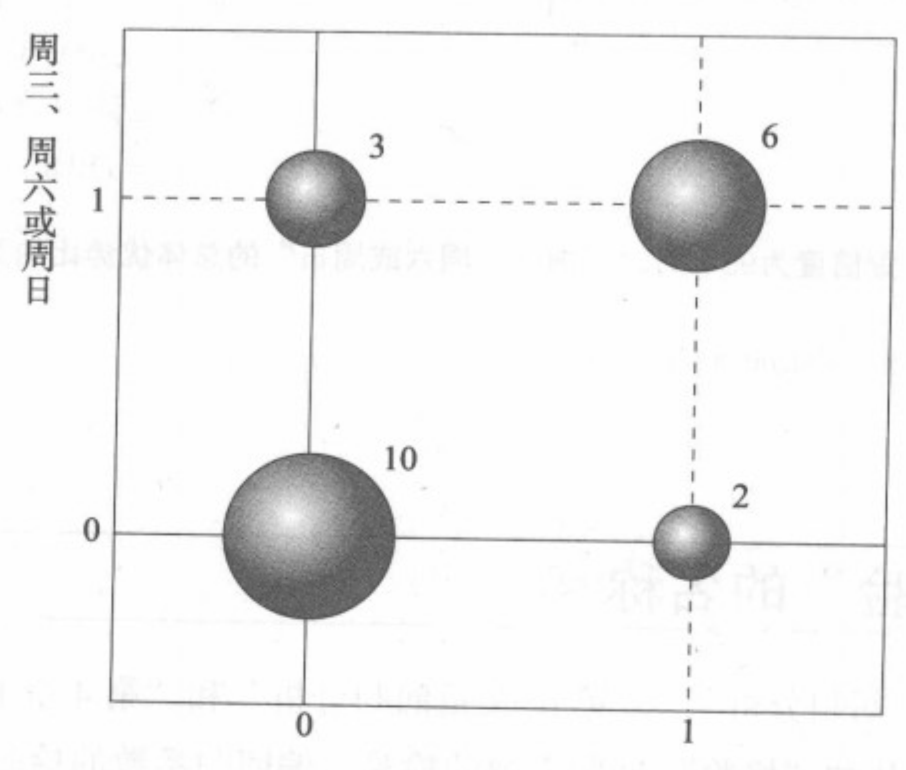
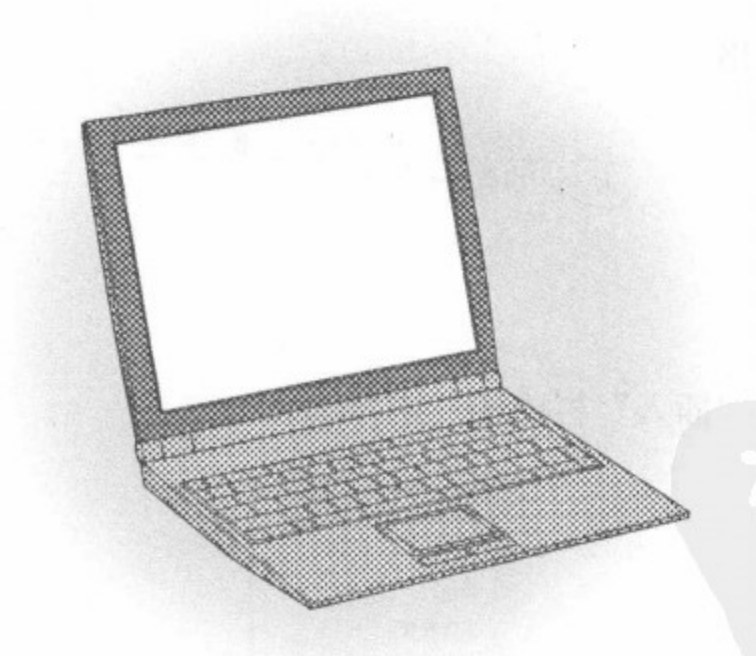


图4.3　同第165页中的图相对应的Bubble Chart

在“周三、周六或周日”卖出的诺伦特供蛋糕，以及“周三、周六或周日”以外没有卖出的诺伦特供蛋糕，就变得一目了然了。

◆附 录◆

用 Excel 算算看



附录中所用数据已经传送到网站 http://www.okbook.co.cn/，名为“数据-回归 .xls”的文件夹中，请您下载使用。

这里将对以下内容进行讲解

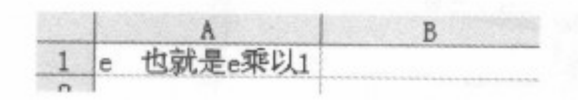
1. 自然对数的底
2. 指数函数
3. 自然对数函数
4. 矩阵的乘法
5. 逆矩阵
6. χ^2 分布的横轴坐标
7. χ^2 分布的概率
8. F 分布的横轴坐标
9. F 分布的概率
10. （重）回归分析的（偏）回归系数
11. Logistic 回归方程的回归系数

1. 自然对数的底

所用数据见第 19 页，均收录在“自然对数的底”表单中。

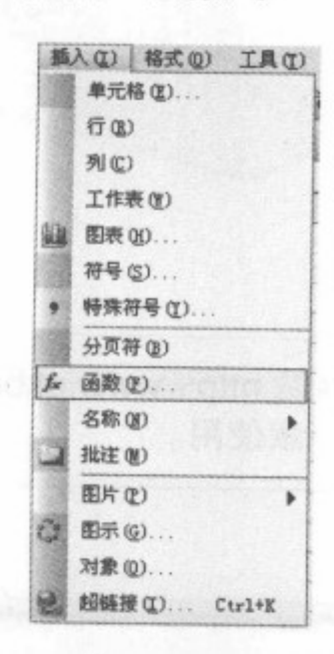
步骤 1

选择“B1”单元格。



步骤 2

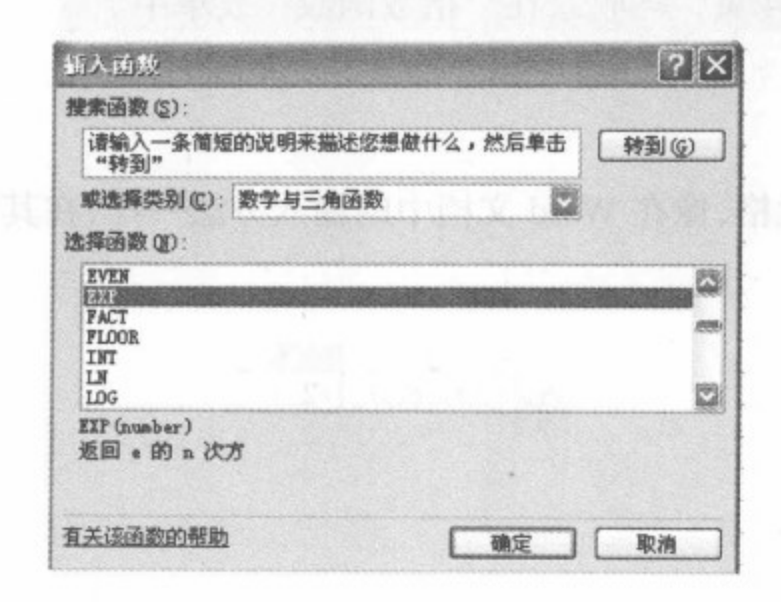
选择菜单栏中的“插入”栏内的“函数”。





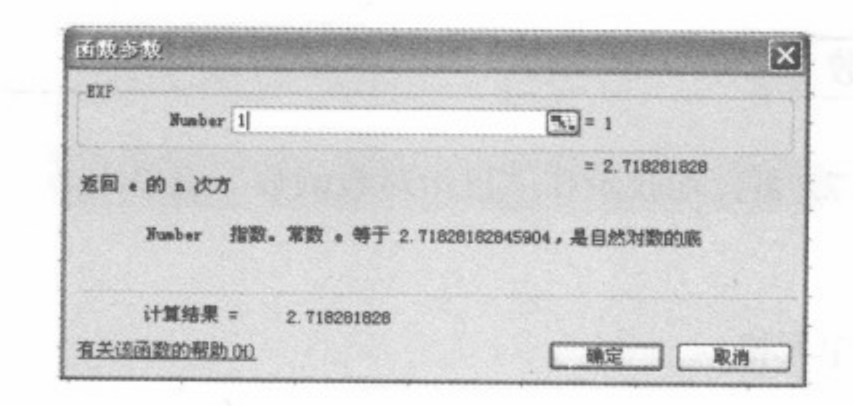
步骤 3

在“选择类别”中选择“数学与三角函数”，在“选择函数”中选择“EXP”。



步骤 4

直接输入 1 后，点击“确定”。



步骤 5

计算完毕。

	A	B
1	e 也就是e乘以1	2.718282

2. 指数函数

所用数据见第 14 页，均收录在“指数函数”表单中。

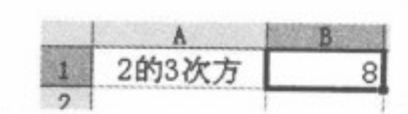
步骤 1

选择“B1”单元格，像在 Word 文档中的输入方法一样，在其中输入“=2^3”后，按下“Enter”键。

	A	B
1	2的3次方	=2^3
2		

步骤 2

计算完毕！

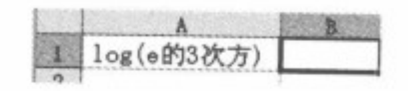


3. 自然对数函数

所用数据见第 22 页，均收录在“自然对数函数”表单中。

步骤 1

选择“B1”单元格。



步骤 2

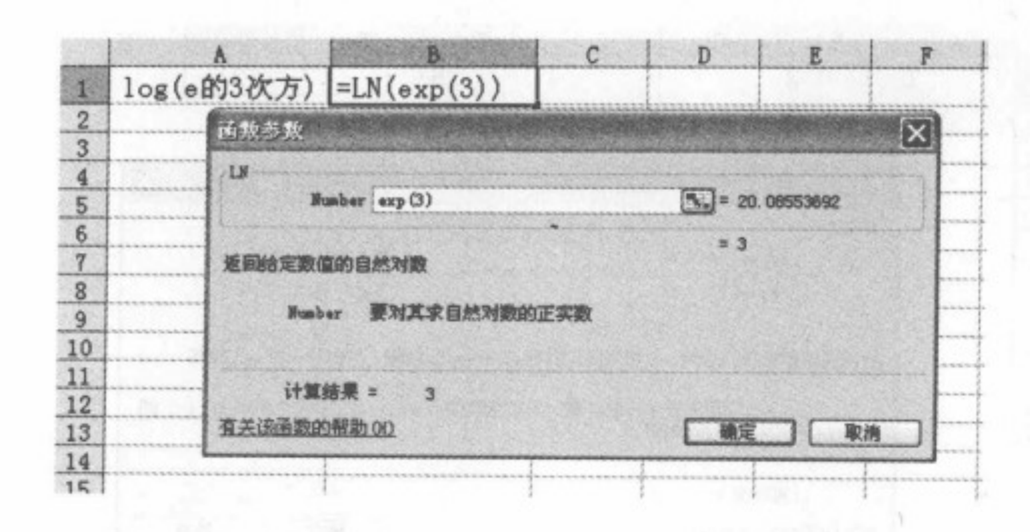
选择菜单栏中的“插入”栏内的“函数”。

步骤 3

在“选择类别”中，选择“数学与三角函数”，在“选择函数”中选择“LN”。

步骤 4

直接输入“exp（3）”后，点击“确定”。



步骤 5

计算完毕！

	A	B
1	log(e的3次方)	3
2		

4. 矩阵的乘法

所用数据见第 41 页，均收录在“矩阵的乘法”表单中。

步骤 1

选择“G1”单元格。

	A	B	C	D	E	F	G
1	1	2		4	5		
2	3	4		-2	4		

步骤 2

选择菜单栏中的“插入”栏内的“函数”。

步骤 3

在“选择类别”中选择“数学与三角函数”,在“选择函数”中选择“MMULT”。

步骤 4

选择下图所示的范围，点击“确定”。

	A	B	C	D	E	F	G	H
1	1	2		4	5		: E2)	
2	3	4		-2	4			

函数参数

MMULT

Array1 A1 : B2 = {1,2;3,4}

Array2 D1 : E2 = {4,5;-2,4}

= {0,13;4,31}

返回两数组矩阵的乘积，结果矩阵的行数与 Array1 相等，列数与 Array2 相等

Array2 用于乘积计算的第一个数组数值，Array1 的列数应该与 Array2 的行数相等

计算结果 = 0

有关该函数的帮助(H)

确定 取消

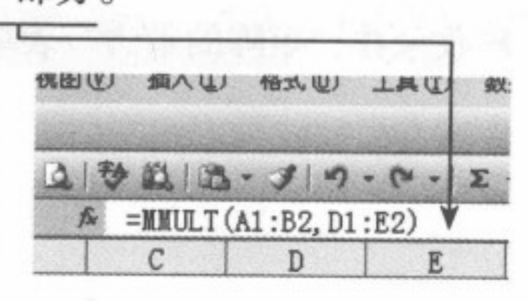
步骤 5

以“G1”单元格为起点，按照下图选择从“G1”到“H2”。

	A	B	C	D	E	F	G	H
1	1	2		4	5		0	
2	3	4		-2	4			

步骤 6

点击数学公式栏中的这一部分。



步骤 7

一起按下“Shift”键和“Ctrl”键，同时再按“Enter”键。

步骤 8

计算完毕！

	A	B	C	D	E	F	G	H
1	1	2		4	5		0	13
2	3	4		-2	4		4	31

5. 逆矩阵

所用数据见第 44 页，均收录在“逆矩阵”表单中。

步骤 1

选择“D1”单元格。

	A	B	C	D
1	1	2		
2	3	4		

步骤 2

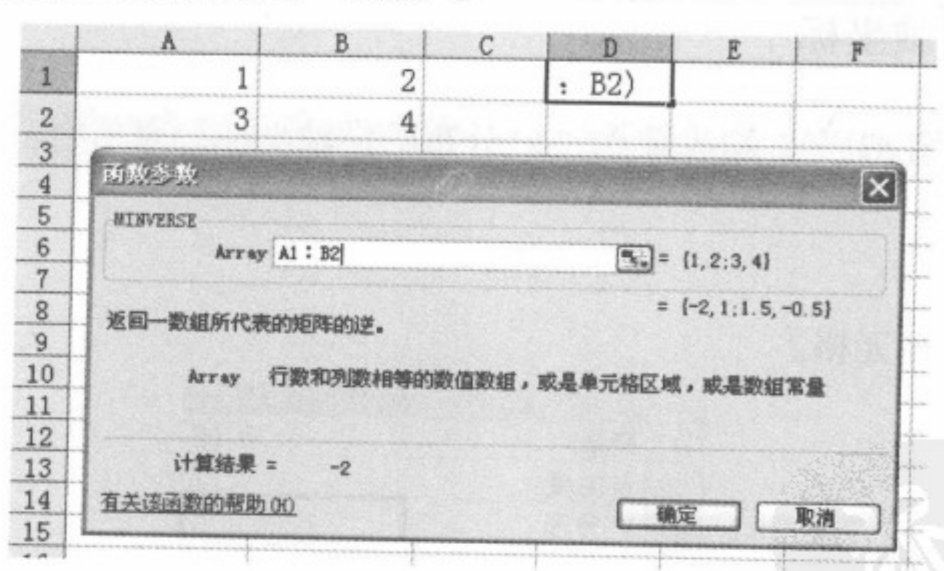
选择菜单栏中的“插入”栏内的“函数”。

步骤 3

在“选择类别”中选择“数学与三角函数”,在“选择函数”中选择“MINVERSE”。

步骤 4

选择下图所示的范围，点击“确定”。



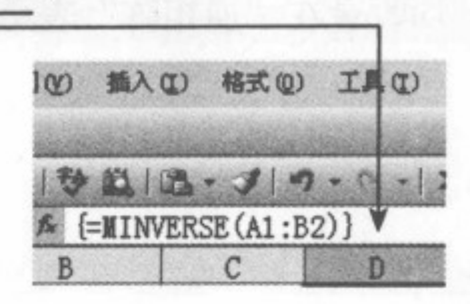
步骤 5

以“D1”单元格为起点，按照下图选择从“D1”到“E2”。

	A	B	C	D	E
1	1	2		−2	
2	3	4			

步骤 6

点击数学公式栏中的这一部分。



步骤 7

一起按下“Shift”键和“Ctrl”键，同时再按“Enter”键。

步骤 8

计算完毕！

	A	B	C	D	E
1	1	2		-2	1
2	3	4		1.5	-0.5
3					

6. χ^2 分布的横轴坐标

所用数据见第 51 页，均收录在“χ^2 分布的横轴坐标”表单中。

步骤 1

选择“B3”单元格。

	A	B
1	概率	0.05
2	自由度	2
3	χ^2分布	
4		

步骤 2

选择菜单栏中的“插入”栏内的“函数”。

步骤 3

在“选择类别”中选择“统计”，在“选择函数”中选择“CHIINV”。

步骤 4

选择“B1”和“B2”单元格，点击“确定”。

	A	B	C	D	E
1	概率	0.05			
2	自由度	2			
3	χ^2分布	[NV(B1, B2)			

函数参数

CHIINV

Probability B1 = 0.05

Deg_freedom B2 = 2

= 5.991464547

返回具有给定概率的收尾 χ2 分布的区间点

Deg_freedom 自由度，介于 1 与 10^10 之间，不含 10^10

计算结果 = 5.991464547

有关该函数的帮助(H) 确定 取消

步骤 5

计算完毕！

	A	B
1	概率	0.05
2	自由度	2
3	χ^2分布	5.991464547

7. χ^2分布的概率

所用数据见第 175 页，均收录在“χ^2分布的概率”表单中。

步骤 1

选择“B3”单元格。

	A	B
1	χ^2分布	10.1
2	自由度	2
3	概率	

步骤 2

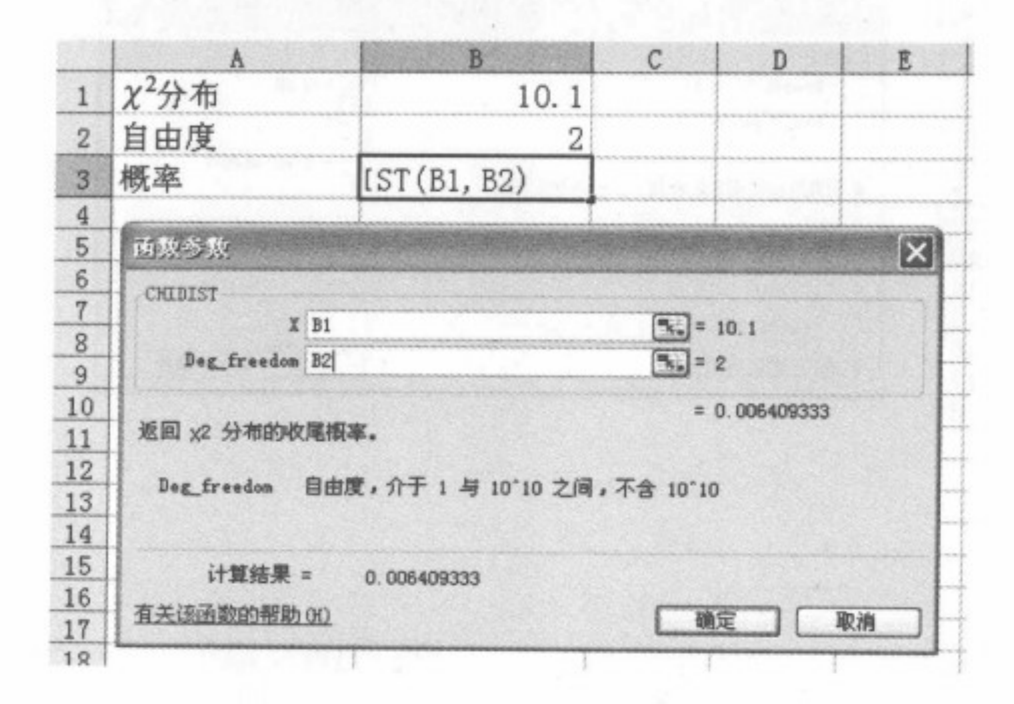
选择菜单栏中的“插入”栏内的“函数”。

步骤 3

在“选择类别”中选择“统计”，在“选择函数”中选择“CHIDIST”。

步骤 4

选择“B1”和“B2”单元格，点击“确定”。



步骤 5

计算完毕！

	A	B
1	χ^2分布	10.1
2	自由度	2
3	概率	0.006409

8. *F*分布的横轴坐标

所用数据见第 54 页，均收录在“*F*分布的横轴坐标”表单中。

步骤 1

选择“B4”单元格。

	A	B
1	概率	0.05
2	第1自由度	1
3	第2自由度	12
4	*F*	

步骤 2

选择菜单栏中的“插入”栏内的“函数”。

步骤 3

在“选择类别”中选择“统计”，在“选择函数”中选择“FINV”。

步骤 4

选择“B1”、“B2”和“B3”单元格，点击“确定”。

	A	B	C	D	E	F
1	概率	0.05				
2	第1自由度	1				
3	第2自由度	12				
4	*F*	B2, B3)				

函数参数

FINV

Probability B1 = 0.05

Deg_freedom1 B2 = 1

Deg_freedom2 B3 = 12

= 4.747225336

返回 F 概率分布的逆函数值，如果 p = FDIST(x,...)，那么 FINV(p,....) = x。

Deg_freedom2 分母的自由度，介于 1 与 10^10 之间，不包含 10^10

计算结果 = 4.747225336

有关该函数的帮助(H) 确定 取消

步骤 5

计算完毕！

	A	B
1	概率	0.05
2	第1自由度	1
3	第2自由度	12
4	*F*	4.747225

9. *F* 分布的概率

所用数据见第 84 页，均收录在“*F* 分布的概率”表单中。

步骤 1

选择“B4”单元格。

	A	B
1	*F*	55.6
2	第1自由度	1
3	第2自由度	12
4	概率	

步骤 2

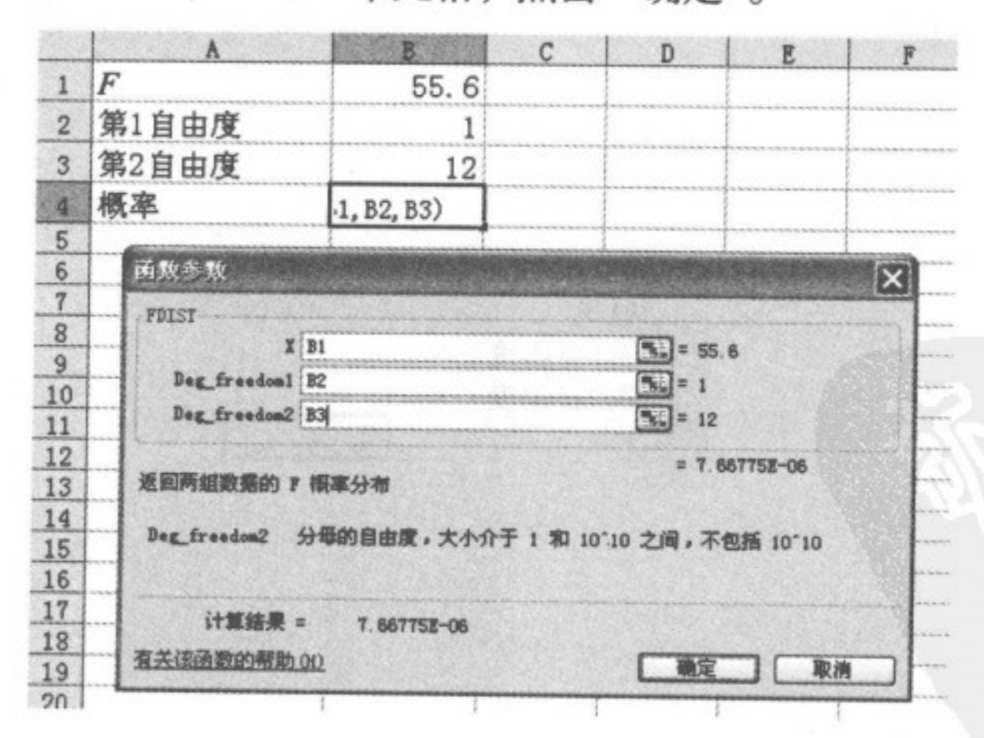
选择菜单栏中的“插入”栏内的“函数”。

步骤 3

在“选择类别”中选择“统计”，在“选择函数”中选择“FDIST”。

步骤 4

选择“B1”、“B2”和“B3”单元格，点击“确定”。



步骤 5

计算完毕！

	A	B
1	F	55.6
2	第1自由度	1
3	第2自由度	12
4	概率	7.67E-06
5		

"7.67E-06"是 Excel 中的书写格式，实际上是"7.67×10^{-6}"。

10.（重）回归分析的（偏）回归系数

所用数据见第 107 页，均收录在"（重）回归分析的（偏）回归系数"表单中。

步骤 1

选择"G2"单元格。

	A	B	C	D	E	F	G	H	I
1		店铺面积	距离最近车站的距离	月营业额			距离最近车站的距离	店铺面积	（常数项）
2	梦之丘总店	10	80	469		偏回归系数			
3	寺井站大厦店	8	0	366					
4	曾根店	8	200	371					
5	桥本大街店	5	200	208					
6	桔梗町店	7	300	246					
7	邮政局前店	8	230	297					
8	水道町站前店	7	40	363					
9	六条站大厦店	9	0	436					
10	若叶川店	6	330	198					
11	美里店	9	180	364					
12									

步骤 2

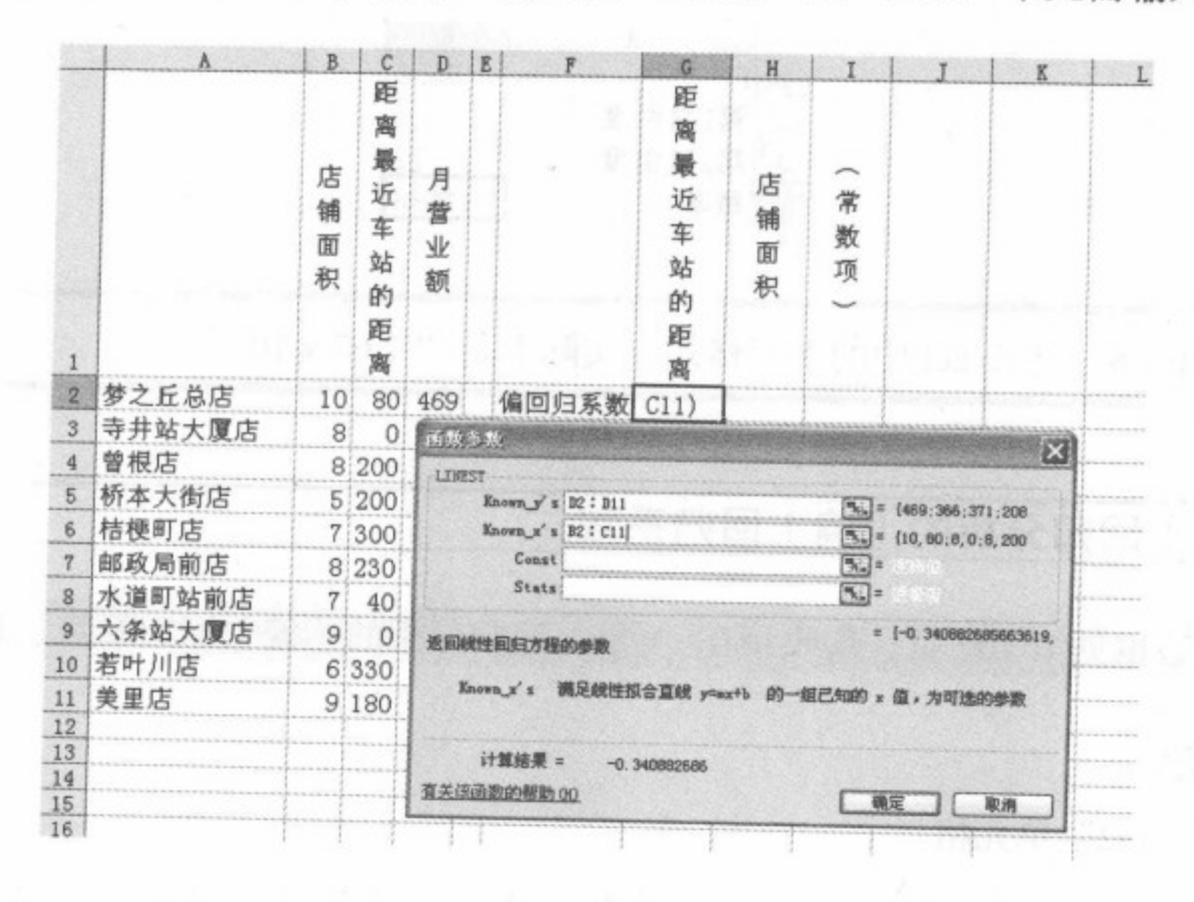
选择菜单栏中的"插入"栏内的"函数"。

步骤 3

在"选择类别"中，选择"统计"，在"选择函数"中，选择"LINEST"。

步骤 4

选择下图所示的范围，点击“确定”。“Const”和“Stats”中无需输入任何值。



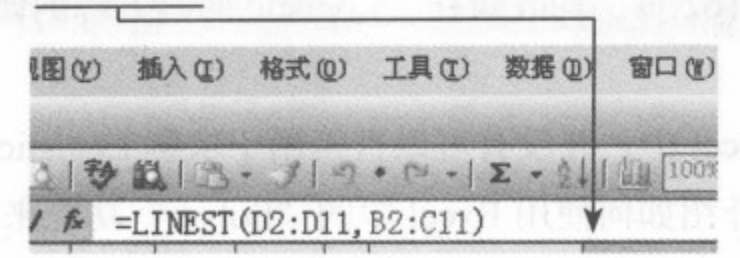
步骤 5

以“G2”单元格为起点，按照下图选择从“G2”到“I2”



步骤 6

点击数学公式栏中的这一部分。



步骤 7

一起按下“Shift”键和“Ctrl”键，同时再按“Enter”键

步骤 8

计算完毕！

G	H	I
距离最近车站的距离	店铺面积	（常数项）
-0.34	41.51	65.32

在“LINEST”函数中，是按照顺序求解（偏）回归系数的。从左到右分别为 $a_p, \cdots, a_2, a_1, b$。

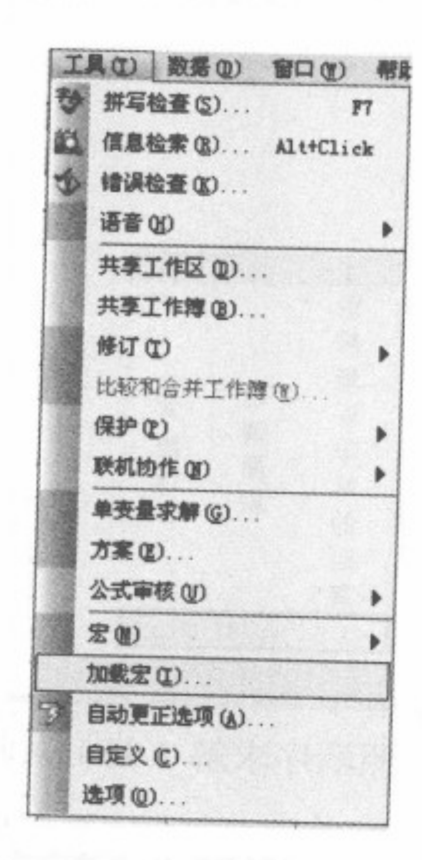
11. Logistic 回归方程的回归系数

所用数据见第 162 页，均收录在“Logistic 回归方程的回归系数”表单中。

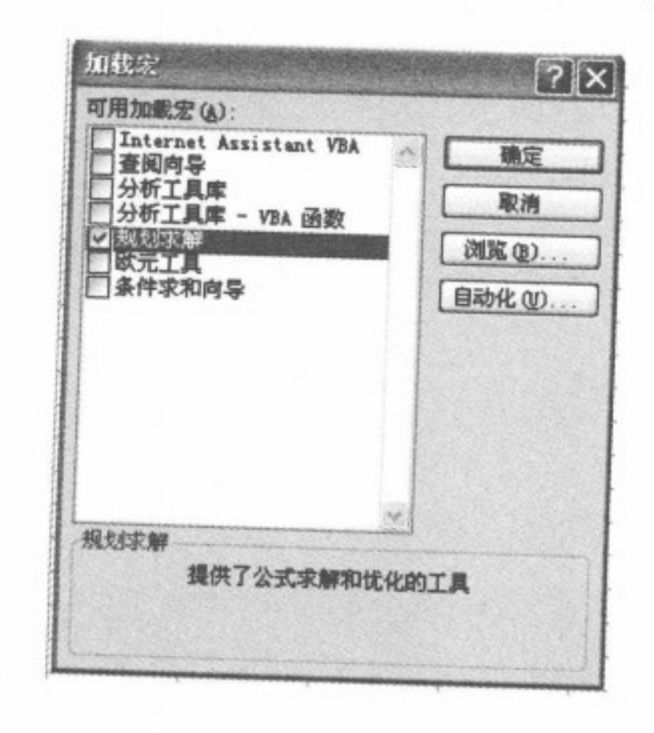
很遗憾，在 Excel 中，并没有可以直接用于求解 Logistic 回归方程的回归系数函数。因此，本节将介绍如何使用 Excel 中的“Solver”功能来求解回归系数。

Solver

① 选择菜单栏中的“工具”栏内的“加载宏”。



② 选择“Solver add in”(求解器加载宏)，点击“确定”。





③ 如果出现“需要 Excel 的安装光盘”等信息，那么请按指示操作。

按照以上步骤操作以后，我们便可以使用这一功能了。

我们可以将第 162 页的数据输入 Excel 表格。如果读者朋友想要求解本节示例以外的 Logistic 回归方程的回归函数，那就要根据实际情况输入相应的 Excel 函数。

步骤 1

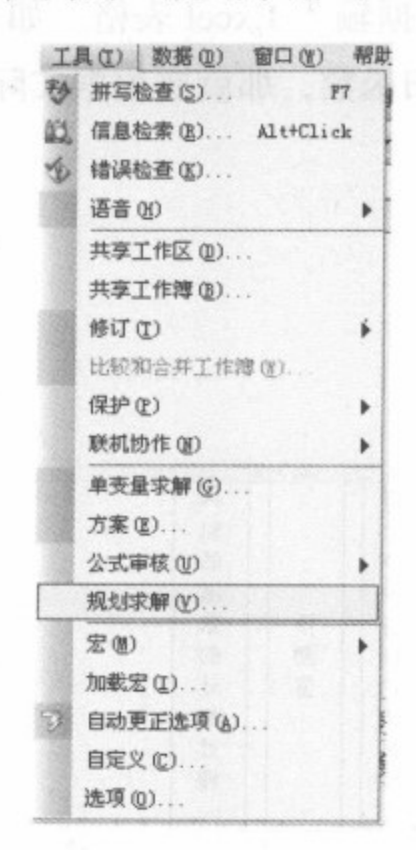
选择“L3”单元格

	A	B	C	D	E	F	G	H	I	J	K	L
1			周三、周六或周日	最高温度	诺伦特供蛋糕的销售情况		预测值		（似然函数的计算过程）			
2	5日	星期一	0	28	1		0.50		0.50		似然函数	4.77E-07
3	6日	星期二	0	24	0		0.50		0.50		对数似然函数	-14.5561
4	7日	星期三	1	26	0		0.50		0.50			
5	8日	星期四	0	24	0		0.50		0.50		a1	
6	9日	星期五	0	23	0		0.50		0.50		a2	
7	10日	星期六	1	28	1		0.50		0.50		b	
8	11日	星期日	1	24	0		0.50		0.50			
9	12日	星期一	0	26	1		0.50		0.50		因变量的值是[1]的个体个数	8
10	13日	星期二	0	25	0		0.50		0.50		因变量的值是[0]的个体个数	13
11	14日	星期三	1	28	1		0.50		0.50		判定系数	-0.04307
12	15日	星期四	0	21	0		0.50		0.50			



步骤 2

选择菜单栏中的“工具”栏内的“规划求解”。



步骤 3

按照下图进行设置，然后点击“求解”。

	A	B	C	D	E	F	G	H	I	J	K	L
1			周三、周六或周日	最高温度	诺伦特供蛋糕的销售情况		预测值		（似然函数的计算过程）			
2	5日	星期一	0	28	1		0.50		0.50		似然函数	4.77E-07
3	6日	星期二	0	24								-14.5561
4	7日	星期三	1	26								
5	8日	星期四	0	24								
6	9日	星期五	0	23								
7	10日	星期六	1	28								
8	11日	星期日	1	24								
9	12日	星期一	0	26								8
10	13日	星期二	0	25								13
11	14日	星期三	1	28								-0.04307
12	15日	星期四	0	21								

规划求解参数

设置目标单元格(E): $L$3

等于: ◉ 最大值(M) ○ 最小值(N) ○ 值为(V) 0

可变单元格(B): $L$5:$L$7 推测(G)

约束(U):

添加(A) 更改(C) 删除(D)

求解(S) 关闭 选项(O) 全部重设(R) 帮助(H)

步骤 4

点击“确定”。

规划求解结果

规划求解收敛于当前的解，可满足所有的约束。

报告(R)
运算结果报告
敏感性报告
极限值报告

⊙ 保存规划求解结果(K)
○ 恢复为原值(O)

确定 取消 保存方案(S)... 帮助(H)

步骤 5

计算完毕！

	A	B	C	D	E	F	G	H	I	J	K	L
1			周三、周六或周日	最高温度	诺伦特供蛋糕的销售情况		预测值		（似然函数的计算过程）			
2	5日	星期一	0	28	1		0.50		0.50		似然函数	4.77E-07
3	6日	星期二	0	24	0		0.50		0.50		对数似然函数	-14.5561
4	7日	星期三	1	26	0		0.50		0.50			
5	8日	星期四	0	24	0		0.50		0.50		a1	
6	9日	星期五	0	23	0		0.50		0.50		a2	
7	10日	星期六	1	28	1		0.50		0.50		b	
8	11日	星期日	1	24	0		0.50		0.50			
9	12日	星期一	0	26	1		0.50		0.50		因变量的值是[1]的个体个数	8
10	13日	星期二	0	25	0		0.50		0.50		因变量的值是[0]的个体个数	13
11	14日	星期三	1	28	1		0.50		0.50		判定系数	-0.04307
12	15日	星期四	0	21	0		0.50		0.50			

在第 109 页中我们讲过，用最小二乘法求解重回归方程的偏回归系数。在“数据 - 回归 .xls” 中，为了练习使用 Solver 功能来求解第 3 章例题中的偏回归系数，我们还准备了“重回归方程 solvr”这样一张工作表。为了进一步熟悉 Solver 功能，同时也为了体会最小二乘法的思想，请您务必要接受这一挑战。



◆ 参考文献 ◆

・市原清志『バイオサイエンスの統計学』（南江堂）1990
・内田治『すぐわかるEXCELによる回帰分析』（東京図書）2002
・内田治 / 菅民郎 / 高橋信『文系にもよくわかる多変量解析』（東京図書）2005
・菅民郎『多変量解析の実践（上）』（現代数学社）1993
・鈴木武 / 山田作太郎『数理統計学―基礎から学ぶデータ解析―』（内田老鶴圃）1996
・高橋信『Excelで学ぶコレスポンデンス分析』（オーム社）2005
・高橋信『マンガでわかる統計学』（オーム社）2004
・高橋善弥太『医者のためのロジスチック・Cox回帰入門』（日本医学館）1995
・丹後俊郎 / 山岡和枝 / 高木晴良『ロジスティック回帰分析』（朝倉書店）1996
・豊田秀樹『調査法講義』（朝倉書店）1998
・永田靖『統計的方法のしくみ』（日科技連）1996
・永田靖 / 棟近雅彦『多変量解析法入門』（サイエンス社）2001
・浜田知久馬『学会・論文発表のための統計学』（真興交易医書出版部）1999


请问诺伦特供蛋糕……
咦？
要买吗？
呵呵

科龙图书读者意见反馈表

书　　名 ________________________________

个人资料

姓　　名：__________ 年　　龄：__________ 联系电话：__________

专　　业：__________ 学　　历：__________ 所从事行业：__________

通信地址：______________________________ 邮　　编：__________

E-mail：______________________________

宝贵意见

◆ 您能接受的此类图书的定价

20 元以内□　30 元以内□　50 元以内□　100 元以内□　均可接受□

◆ 您购本书的主要原因有(可多选)

学习参考□　教材□　业务需要□　其他__________

◆ 您认为本书需要改进的地方(或者您未来的需要)

____

____

◆ 您读过的好书(或者对您有帮助的图书)

____

____

◆ 您希望看到哪些方面的新图书

____

____

◆ 您对我社的其他建议

____

____

谢谢您关注本书！您的建议和意见将成为我们进一步提高工作的重要参考。我社承诺对读者信息予以保密，仅用于图书质量改进和向读者快递新书信息工作。对于已经购买我社图书并回执本“科龙图书读者意见反馈表”的读者，我们将为您建立服务档案，并定期给您发送我社的出版资讯或目录；同时将定期抽取幸运读者，赠送我社出版的新书。如果您发现本书的内容有个别错误或纰漏，烦请另附勘误表。

回执地址：北京市朝阳区华严北里 11 号楼 3 层

科学出版社东方科龙图文有限公司电工电子编辑部(收)



邮编：100029

（N-0355.0101）
责任编辑：唐　璐　赵丽艳
责任制作：董立颖　魏　谨
封面制作：铭轩堂设计：13671110894 铭志辉鸿 轩雅静氣

用漫画这种形式讲数学、物理和统计学，十分有利于在广大青少年中普及科学知识。

周恩来、邓颖超秘书，周恩来邓颖超纪念馆顾问
中日友好协会理事，《数理天地》顾问，全国政协原副秘书长



用漫画和说故事的形式讲数学，使面貌冷峻的数学变得亲切、生动、有趣，使学习数学变得容易，这对于提高全民的数学水平无疑是功德无量的事。

《数理天地》杂志社　社长 总编
"希望杯"全国数学邀请赛组委会 命题委员会主任

周国镇

用漫画的形式，讲解日常生活中的数学、物理知识，更能让大家感受到数学殿堂的奥妙与乐趣。

《光明日报》原副总编辑
中华炎黄文化研究会 常务副会长

鲁谆

科学漫画是帮助学习文科的人们用形象思维的方式掌握自然科学的金钥匙。

中国人民大学外语学院日语专业 主任
大学日语教学研究会 会长

成同社

在日本留学的时候，我在电车上几乎每次都能看到很多年轻的白领看这套图书，经济实惠、图文并茂、浅显易懂，相信这套图书的中文版也一定会成为白领们的手中爱物。

大连理工大学 能源与动力学院 博士 副教授

我非常希望能够在书店里看到这样的书：有人物形象、有卡通图、有故事情节，当然最重要的还有深厚的理工科底蕴。我想这样的书一定可以大大提升孩子们的学习兴趣，降低他们对于高深的理工科知识的恐惧感。

北京启明星培训学校 校长

书中的数学知识浅显实用，漫画故事的形式使知识贴近生活，概念更容易理解。

北京大学 数学科学学院 博士

张磊



定价：29.80元

科学出版社 东方科龙
http://www.okbook.com.cn
zhaoliyan@mail.sciencep.com